

Nanotech is not so scary

Given recent rumblings from opinion-formers, researchers working on the science of the incredibly small should exert more effort on putting the risks posed by their work into the proper perspective.

Is nanotechnology inherently dangerous? Calls for regulation of this rapidly developing and diverse discipline seem to imply that it is. Only last week, for instance, Britain's Better Regulation Task Force, in a report on the regulation of scientific research, urged the UK government "to demonstrate it has clear policies in place to ensure the safety of individuals, animals and the environment", in the face of developments in the field.

The task force's recommendations are unremarkable, even anodyne: pleas for openness, informed public debate, and foresight over potential risks. More noteworthy is the fact that nanotech was singled out for attention in a report that otherwise dealt with established flashpoints for public concern, such as genetically modified crops and human embryonic stem cells. The task force's fear seems to be that nanotechnology might soon become the topic of public unease — and that the resulting debate will take place in an informational vacuum that will quickly be filled with hot air and hysteria.

This is a valid point. Possible risks of nanotech are already under a media spotlight, but the debate is informed largely by science fiction. Literally so: Michael Crichton's new thriller *Prey*, in which self-replicating nanoscale robots run amok, has for weeks been riding high in the best-seller lists. When the inevitable movie appears, a nanotech Armageddon will become popular fare.

Stories told by scientists are always fair game for novelists looking for a racy plot device. And this is an old story, related most famously in 1986 by Eric Drexler in his book *Engines of Creation*. This visionary account of what nanotech might offer to medicine and other human endeavours also warned of a darker side: the possibility that 'nanobots' created to build structures and materials, including copies of themselves, atom by atom, might start replicating endlessly.

These rogue nanobots, Drexler envisaged, would then pull apart everything around them and use it to build copies of themselves, ultimately turning the world into a 'grey goo'. This has become a favourite scenario, and was revived most prominently by Bill Joy, chief scientist at Sun Microsystems, in an article in *Wired* magazine in April 2000.

But the way in which nanotech has developed since Drexler's words of warning makes the grey-goo nightmare an unlikely prospect. No one is considering building nanobots that replicate by manipulating atoms one at a time, and several leading figures in nanotech research argue that the whole idea is unfeasible. Self-replication is certainly one focus of nanotechnology, but it is hard to see how any of the avenues currently being explored could even become autonomous, let alone get out of control. Anyone wishing to make mischief with self-replicators would start with cells or viruses, not hypothetical nanobots.

There may well be dangers in nanotechnology, as in any emerging area of research. We should certainly look at the potential toxicity of nanoparticles being touted as medical diagnostic tools. But nanotechnology is a diverse field, united only by the factor of scale. So it is not even clear how one would go about regulating nanotech in a manner unique to the discipline.

It is tempting to conclude that nanotech would not be subjected to suspicious scrutiny at all were it not for the enduring but outdated image of grey goo. But that is all the more reason why the Better Regulation Task Force's proposal for good public information should be heeded. It would do no one any favours to leave public opinion, and government regulations, to be shaped by the writers of science-fiction thrillers. ■

Fair play for trial balloons

Ultra-long-duration ballooning has much to offer astronomers — but only if NASA nurtures this fledgling technology.

Consumer gadgets, such as cameras and computers, tend to advance in quality even as their price comes down. Unfortunately for space scientists, rockets aren't like that. The launch vehicles used for today's spacecraft are fundamentally little different from the glorified guided missiles that delivered the first Sputniks into orbit almost half a century ago. And they are still fantastically expensive, forcing NASA and other space agencies to spend billions of dollars each year on launch fees.

Worse, rockets still fail with disturbing regularity — just ask the scientists working on the European Space Agency's Rosetta comet mission, who are now rearranging their professional lives and scratching around for further funding because a glitch in the Ariane 5 'heavy lifter' has grounded the spacecraft (see page 301).

If scientific instruments are to land on a comet, there is no choice but to rely on today's imperfect rocket science. But for astronomers who merely need to loft their telescopes above the distorting effects of the Earth's atmosphere, a new capability called ultra-long-duration ballooning offers a cheap alternative. While not quite offering a ride

into space, a new generation of balloons could allow telescopes to soar above 99% of the atmosphere for more than 100 days at a time — and for a fraction of the cost of launching a space telescope.

The technology has an important qualifying test this month (see page 308). But even though they stand on the brink of an exciting new era, enthusiasts for balloon-based astronomy have nagging worries about NASA's commitment to their cause. They are frustrated, for instance, by the fact that the space agency has never opened up a dedicated line of funding for balloon-based astronomy. And they now fear that, if the ultra-long-duration programme suffers a failure or two, NASA's managers will swiftly bail out.

In part, NASA's limited enthusiasm for ballooning is a reflection of the political environment in which the agency operates: the small firms that manufacture scientific balloons lack the lobbying clout of aerospace giants such as Boeing and Lockheed Martin. But in the current harsh economic climate, NASA should place a high premium on the potential cost-effectiveness of balloon-based astronomy, and remember that rockets are still far from perfect after all these years. ■

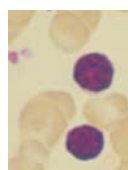




Divisive issue
Resignations add
up to trouble for
maths institute
p303



Switching on
Tokyo University
reforms let students
change course
p304



Second setback
Another child in
French gene-therapy
trial has leukaemia
p305



Landmark
Nature's founder
honoured by
heritage body
p307

Failed mission launch prompts ESA to reconsider comet target

Declan Butler, Paris

STUNNING SPACECRAFT SEEKS COMET. Should be small and highly active, and reachable by launch from Earth within the next 2.5 years. Rendezvous must require no major engineering works, and incur few additional costs. Technically risky options need not apply.

Europe's comet scientists are this week struggling to determine a new route — and probably a new target — for Rosetta, the world's largest and most ambitious comet mission, whose postponed launch last week saw it miss the January window that would have allowed the probe to land on the comet Wirtanen in 2010 (see *Nature* 421, 198; 2003).

The scientists now face a formidable challenge if they are to realize Rosetta's exacting goals. The truck-sized craft must be lobbed into the outer reaches of the Solar System, and accelerated to the 9 kilometres per second needed to catch up with its tiny target, a dirty snowball barely a few kilometres across. The probe then has to orbit the comet, plant a lander on it, and hang on while the comet arcs around the Sun.

The sooner Rosetta is launched, the less the postponement will cost the European Space Agency (ESA) — most costs are derived from retaining the design and development teams until the launch. But most of the earliest launch possibilities for Rosetta carry major technical risks.

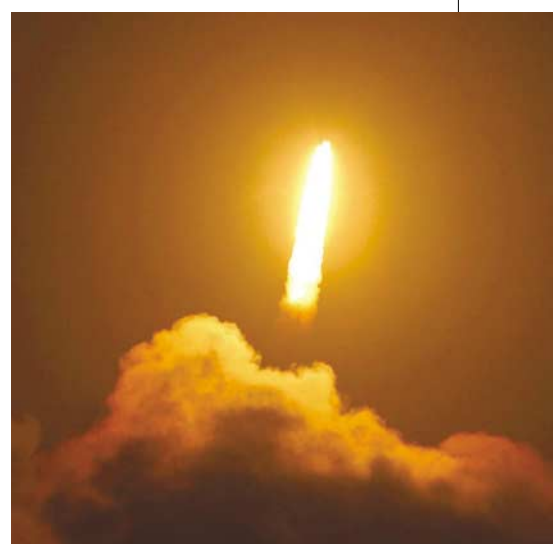
Two dates later this year would use Venus, instead of Mars, to provide a gravitational 'catapult' that would enable Rosetta to catch up with its original target, Wirtanen. But these options are unlikely to be selected, according to Gerhard Schwehm, who leads the core Rosetta project team from ESA's European Space Research and Technology Centre (ESTEC) in Noordwijk, the Netherlands. The probe was designed for cold space, rather than the cauldron of the inner Solar System. "The heat load is too high for some of the spacecraft's subsystems to survive," he says. David Southwood, ESA's director of science agrees: "You can't just install a fan," he quips.

A January 2004 launch window, on the other hand, would retain both the Mars pass and the 2010 Wirtanen rendezvous. This

particular flight plan would call for a rocket launch velocity that is higher than previously planned, however. That would leave mission planners facing the unenticing prospect of using the Ariane 5 'heavy lifter' that blew up in December — the incident that created the problem in the first place. Schwehm prefers to stick with the original launcher, especially as there is no guarantee that the souped-up model will be flying safely by January 2004.

That leaves scientists with what might be their best bet: shooting for another comet from the Jupiter family, so called because its members' orbits cross that of Jupiter and are influenced by the planet's gravitational pull. The likely target will be in a phase and orbit where it is producing a lot of hot gas and dust, says Schwehm.

This will involve reviewing reams of existing observations, and making new ones. Problems with the various options include



An explosion of the Ariane launcher caused the Rosetta comet probe to miss its launch window.

J. AMIET/AFP

Report backs military axis on space

Declan Butler, Paris

Europe should consider a sweeping reorganization of its space programme to integrate its civilian and military endeavours, argues a French report released last Friday.

If approved, the reorganization would be the largest since the European Space Agency (ESA) was founded in 1975.

The report, from a commission headed by Roger-Maurice Bonnet, who led ESA science for a decade before retiring in 2001, speaks of a "major crisis" in European space policy. It calls for the revival by 2005 of a high-level European Space Conference to spearhead the programme's revival.

Contentiously, the report also says that the programme should be led by military requirements — albeit while embracing the interests of science, agriculture, navigation, telecommunications, environment and technology development. The United States spends six times more than Europe on its space programmes, it notes.

"We think that the major effort of the United States in military space is a lesson for Europe," says Bonnet, adding that the transformation of ESA into a joint military-civilian agency is only one possible route to building a stronger programme. The report also advocates government intervention to consolidate Europe's fragmented aerospace and defence industries.

The French government is likely to back the report when it responds to it in March — but some other governments, less taken by the idea of competing with the United States in this sphere, may not be so enthusiastic.

Nevertheless, the explosion of the Ariane V rocket late last year points to the need for reform. Arianespace, the rocket's maker, is struggling to remain profitable. France's national space agency, the CNES, the largest in Europe, is in a budgetary crisis, and other national space programmes are stagnant or in decline.

possible design changes needed to the probe, and how the size and surface of the target comet will affect the lander's descent plan.

Rosetta's planners currently seem to favour a spring 2004 launch that would allow targeting of the comet Churyumov–Gerasimenko; other possible windows would allow visits to the comets Howell, Finlay, Schwassmann–Wachmann 2 and Wild-2. Launch dates for these options cannot be firmly defined until scientists at ESA's Operations Centre in Darmstadt, Germany, take ESTEC's proposals and work out the various possible trajectories to get the craft to its potential targets. But these targets might not be reachable until 2015 — a painful delay for those involved in a mission for which planning began in 1993.

Rosetta teams will meet on 13 February to consider possible solutions to the problem, before presenting them to ESA's Science Programme Committee on 26 February. But the crunch will come in May, when ESA will select one of the options — and decide how to pay for it. Southwood estimates that the delay will inflate the mission's £1-billion (US\$1-billion) price tag by a further £50 million–100 million.

ESA may ask its member states to make a one-off contribution to cover this, although with national space budgets under pressure, this may not be forthcoming (see previous page). That could force the agency to make cuts elsewhere in its science programme — possibly including the cancellation of Venus Express, ESA's planned 2005 mission to Venus. ■

Mozambique prime minister tipped for global health post

Declan Butler, Paris

An African official is emerging as the front-runner to succeed Gro Harlem Brundtland as director-general of the World Health Organization (WHO) in May.

If he is elected, Mozambique's prime minister Pascoal Mocumbi will be the first African to head the WHO. The chances that Mocumbi will win a secret ballot of the WHO's 32-strong executive board to be held next week increased after his strongest African rival, former Senegalese health minister Awa Marie Coll-Seck, withdrew from the contest. Mocumbi says that he would resign from his current position if appointed to lead the Geneva-based organization.

Others shortlisted are Julio Frenk Mora, Mexico's health minister; Peter Piot, a former AIDS researcher who now heads the Joint UN Programme on HIV/AIDS (UNAIDS); Jong Wook Lee, a physician from South Korea, who heads the WHO's tuberculosis programme; and former Egyptian health minister Ismail Sallam. The World Health Assembly is expected to endorse the board's choice when it meets in Geneva in May.

The WHO is under strong pressure to end the political horse-trading that traditionally characterizes its elections. This time, *The Lancet* and the New York-based Rockefeller Foundation have organized a campaign to introduce greater openness into what they describe as the "secret and generally unaccountable process" of the election. ■



Pascoal Mocumbi (left) and Peter Piot are in the running to head the World Health Organization.

Critics of the election process say that it reflects the highly politicized workings of many international agencies during the cold war, and should be replaced by a more open and merit-based contest when the next director-general is elected in 2008.

Under pressure to fight an open campaign, candidates initially agreed to take part in a two-hour question-and-answer session on 19 January, set up by non-governmental organizations and the Interactive Health Network's World Health Channel, and broadcast live on the Internet. But several of the candidates pulled out of the session.

Each candidate has, however, taken up *The Lancet's* challenge to provide detailed responses on its website to ten questions about the WHO's future direction. ■

Bushfires annihilate Australian observatory

Carina Dennis, Sydney

One of Australia's premier observatories has been destroyed by bushfires that have raged for days in and around the nation's capital, Canberra. Flames engulfed Mount Stromlo



Gutted: the charred remains of Mount Stromlo Observatory — remarkably, some data survived.

Observatory, which lies at the western outskirts of the city, on 18 January, melting telescope housings, incinerating workshops and destroying precious archives.

"It looks like it has been firebombed," says Brian Schmidt, a research fellow at the Australian National University's astronomy school in Canberra, which operates the observatory. "All our active telescopes have been destroyed."

No astronomers were killed or injured in the fires — which caused four deaths and injured hundreds in the area — but several of their homes near the observatory were destroyed. Preliminary estimates value the observatory's losses at more than A\$20 million (US\$12 million), although Schmidt thinks the final figure could be twice as much.

Many research projects have been cut short or temporarily halted by the fire, including a five-year programme led by

Schmidt to digitally map the southern sky.

The observatory's construction workshops were destroyed in the blaze, including those that were to be used for a A\$6.3-million contract to build an imager for the Gemini South Telescope in Chile. A A\$5-million near-infrared integral field spectrograph for the Gemini North Telescope in Hawaii, which was ready for shipment, was also damaged.

Historical records have been lost, including drawings of astronomical objects, some dating from the nineteenth century. But buildings containing computers and databases were spared. "Our current research data seem to be intact," says Schmidt.

The Australian National University has already given assurances that the observatory will be rebuilt, with resources and facilities made available to researchers to fulfil equipment contracts in the interim. ■

Resignations rock mathematics institute

Geoff Brumfiel, Baltimore

A string of resignations at the Massachusetts-based Clay Mathematics Institute (CMI) has left some top mathematicians uneasy about the future of one of the discipline's most benevolent backers.

Last November, CMI president Arthur Jaffe — a mathematical physicist at Harvard University and recent past president of the American Mathematical Society (AMS) — resigned, with two other members of the four-strong scientific advisory board.

Concerns about the resignations were voiced at the AMS annual meeting in Baltimore, Maryland, on 15–18 January. “I’m very worried,” says Peter Sarnak, a mathematician at both Princeton and New York universities. “For pure mathematics, the CMI has been a godsend. If it disappeared, it would be very serious.”

The CMI was set up in 1998 by Landon Clay, a local financier and philanthropist, with Jaffe as its first president, to fund innovative research and raise the profile of mathematics. It is perhaps best known for offering prizes worth US\$1 million apiece to anyone who could solve seven famous mathematical problems (see *Nature* 405, 383; 2000).

The institute also provided about \$3 million in grants and awards worldwide last year. “The Clay has raised our visibility and provided a significant amount of resources to mathematicians,” says Hyman Bass of the University of Michigan, the current AMS



Arthur Jaffe: walked out of the institute in November.

Seattle-based Microsoft Research.

Jaffe says that he was told in November he would have to leave at the end of 2002, but that no reason was given for the request. In the meantime, he says, the board of directors proposed constraints on his interactions with other members of the science board that made it impossible for him to do his job.

Eric Woodbury, the CMI's chief administrator, maintains that Jaffe's three-year term as CMI president had simply expired. Jaffe disputes this. “To my knowledge, there was no term,” he says. There is no mention of a term in the CMI's by-laws.

The other board members who resigned were Edward Witten, of the Institute for Advanced Study at Princeton, and Alain Connes, of the Collège de France in Paris. Both are recipients of the discipline's highest

honour, the Fields medal. A copy of Connes' resignation letter, obtained by *Nature*, cites Jaffe's removal as his reason for quitting. “The main reason is that I disagree with the dismissal of Arthur Jaffe as president and do not want to assume any responsibility for this move,” Connes wrote.

Woodbury says that the institute is undergoing a normal transition. “It is expected there will be turnover from time to time,” he says. Three new members, including two Fields medal winners, have agreed to join the scientific board, he says. Woodbury expects a new president to be appointed this year. “Our level of programming and research support will be continued,” he says. “And the \$7 million in prizes won't go away.”

Woodbury says that the institute is undergoing a normal transition. “It is expected there will be turnover from time to time,” he says. Three new members, including two Fields medal winners, have agreed to join the scientific board, he says. Woodbury expects a new president to be appointed this year. “Our level of programming and research support will be continued,” he says. “And the \$7 million in prizes won't go away.”

Woodbury says that the institute is undergoing a normal transition. “It is expected there will be turnover from time to time,” he says. Three new members, including two Fields medal winners, have agreed to join the scientific board, he says. Woodbury expects a new president to be appointed this year. “Our level of programming and research support will be continued,” he says. “And the \$7 million in prizes won't go away.”



Researcher uncovers truth behind wartime security slur

Rex Dalton, San Diego

Austrian-born oceanographer Walter Munk has never forgotten the false allegations of Nazi sympathies that prompted the US government to strip him and his mentor of their security clearances six decades ago.

A disclosure of government files has now confirmed the thinness of the case prepared against the two by J. Edgar Hoover, director of the Federal Bureau of Investigation (FBI).

Munk is keen to tell his story to today's scientists — some of whom may also find their loyalties questioned on the basis of their ethnic origins. He has recently published an account of the events (*Oceanography* 15, No. 4, 7; 2002), after obtaining his case files under a Freedom of Information Act request.

Munk and his Norwegian mentor, Harald Sverdrup, lost their clearances in 1942, while working at the Scripps Institution of Oceanography (SIO) in La Jolla, California. The duo was involved in a wave-prediction model that was crucial in aiding the US

Navy's first amphibious landing of the war in North Africa.

Sverdrup, a prominent Norwegian oceanographer who was the SIO's director from 1936 to 1948, never again received full security clearance, although he acted as a government consultant on military projects. He died in Norway in 1957, never knowing why his clearance had been withdrawn.

Munk, now aged 85, won back his military security clearance soon after it was pulled — helped by leading scientists who wrote to the government on behalf of the duo — and is still a senior SIO researcher.

Records held by the Navy and FBI show that the clearances were withdrawn as a result of statements from two now-dead faculty colleagues at Scripps — bacteriologist Claude ZoBell and biochemist Denis Fox — and of some other acquaintances.

Munk and Naomi Oreskes, an Earth-science historian at Scripps who has also written about the affair, believe that the

investigation may have been driven by grudges arising from Sverdrup's leadership of Scripps. There was no evidence of Nazi collaboration or misdeeds in the files, only records of the informants' testimony, possibly based on the fact that the two were foreign and spoke German.

Munk was “amazed” on reading the sum of the evidence against him and Sverdrup, he says. The investigating agencies “deprived the country of needed talents and put the careers of several individuals in jeopardy”, Munk writes in his article.

Oreskes and Ronald Rainger, a historian at Texas Tech University in Lubbock, asserted in an article on the affair (*Stud. Hist. Phil. Mod. Phys.* 31, 309–369; 2000) that the probe helped to create a framework for subsequent government investigations of physicists, including Robert Oppenheimer, engaged in the Manhattan Project to build the atom bomb. “In matters of loyalty, the burden of proof was on the accused, not the accuser,” they say.

Tokyo deans plan flexible future for students

Keiko Kandachi, Tokyo

Undergraduates at the University of Tokyo will have unprecedented freedom to switch between the sciences and the humanities, under a reform package that is due to take effect in three years' time. The move, which will be announced in March, is seen as a harbinger of change in Japan's notoriously inflexible higher-education system.

At the moment, it is difficult for students at Japan's most prestigious university to change their academic course after admittance. But the university's deans recently agreed to reforms that will allow science undergraduates to transfer to law, for example, at the end of their second year of study. Each department will accept a certain percentage of students from any academic track, based on their grades in their first two years.

The idea is to give students enough leeway to do what really interests them, explains Atsushi Koma, the university's vice-president. It is also aimed at meeting a growing demand for graduates who are well-versed in both sciences and the humanities to work in business, law, government and academia.

Under the plan, each department will require students to have attended prerequisite classes before transferring. The university promises to notify students of these requirements early enough for them to make a smooth transfer and graduate within the regular time frame.

In a bid to give students more freedom of

choice, the university will introduce several associated measures, including a lecture series about new developments in the sciences and humanities, to help students to identify their true interests. The university will also introduce a financial-aid plan for outstanding students.

Koma says that in his experience as a chemistry professor, current restrictions stifle students by forcing them to persevere with topics they don't like. He adds that even students that move on to postgraduate work often lack a real passion for their subject area, and so avoid challenging choices of subject for their theses. The reforms will encourage students to take classes in more subjects, so they find out what intrigues them, Koma says.

But Masakata Ogawa, a professor of science and technology education at Kobe University, says that students will need much better career-planning advice than they get at the moment if they are to take advantage of a more flexible course structure.

Students are also voicing scepticism about the change. "Before the university introduces the reforms, I want it to make sure that each professor is more committed to teaching," says one science student, who did not want to be identified. "Some professors seem to focus on research and lack the energy needed to teach undergraduates."

But education experts say that the initiative will probably lead the way for Japan's



Officials at the University of Tokyo hope reforms will ignite some passion in the student body.

K. KANDACHI

other universities. "The innovation at the nation's top institution is likely to become a model for the Japanese university system," says Masamichi Ishii, a senior research fellow at the National Institute of Science and Technology Policy in Tokyo. ■

Transgenic salmon still out in the cold in United States

Hannah Hoag, Washington

Approval for the commercial farming of transgenic salmon in the United States remains at least 18 months off, say officials at Aqua Bounty Farms in Waltham, Massachusetts, the only company that has filed for approval to produce them.

The progress of the application, made by Aqua Bounty Farms in November 1996, has been watched avidly by advocates of transgenic technology — including researchers hoping to develop their own genetically modified fish — who had hoped that the Food and Drug Administration (FDA) would have given its approval by now (see *Nature* 406, 10–12; 2000).

Critics of the technology point to the unknown risks posed by a possible escape of the fish from their farms, perhaps leading to competition with wild fish for food, the introduction of transgenes into wild fish, and even possible collapse of wild fish populations.

Joseph McGonigle, vice-president of

Aqua Bounty Farms, says that he doesn't expect the FDA to approve the application until late 2004. The company plans to submit a second application, for transgenic trout, in about a year's time.

McGonigle says that the delay is placing the company under financial strain. "Money's an issue now," he says. "We've been living on venture capital for years."

But a report issued last week by the non-profit Pew Initiative on Food and Biotechnology questioned the FDA's ability to regulate transgenic fish properly. It says that the FDA lacks the expertise to evaluate environmental risks. The report also calls for greater transparency and public involvement in the review process.

Aqua Bounty's Atlantic salmon carry a gene construct of the growth-hormone gene from Pacific chinook salmon, and a promoter sequence from the ocean pout for the production of antifreeze proteins. They grow twice as fast as most farmed salmon but consume less food. The company is

working with the Harvard Center for Risk Analysis to perform an environmental-impact assessment, which it will use to support its application.

Meanwhile, Chinese researchers have been working with fast-growing transgenic carp since the late 1980s, and Cuba is considering commercialization of tilapia that were first bred there in 1993. The Cuban researchers say it will be three years before they know the regulatory fate of their tilapia.

Mario Estrada, leader of the aquatic-organism project at the Center for Genetic Engineering and Biotechnology in Havana, says that most of the food-safety studies are done, including an experiment in which tilapia were eaten by volunteers.

Researchers elsewhere are working with transgenes that will resist bacterial or viral infection. "Disease is one of the biggest issues in raising fish in aquaculture," says Bill Muir, a professor of genetics at Purdue University in West Lafayette, Indiana. ■

♦ <http://pewagbiotech.org/research/fish>

Second cancer case halts gene-therapy trials

Erika Check, Washington

The world of gene therapy was shaken last year when a child treated in a French trial developed leukaemia. Researchers had pinned their hopes on this being an unfortunate one-off. Now those hopes have been dashed with the emergence of a second, almost identical case that could jeopardize the future of gene therapy.

The latest case centres on a three-year-old boy treated in a gene-therapy trial led by Alain Fischer at the Necker Hospital for Sick Children in Paris. Just under three years ago, the child was cured of severe combined immunodeficiency disease (SCID), a condition that disrupts the development of the immune system. But researchers revealed last week that the boy was diagnosed with leukaemia just days before Christmas.

The news last August that a child in Fischer's trial had developed cancer left nations divided over their regulatory response (see *Nature* 419, 545–546; 2002). This time there was greater accord. Britain has stopped the world's only active SCID trial, and the US Food and Drug Administration (FDA) has suspended more than two dozen similar gene-therapy trials in a variety of diseases — trials it had allowed to continue after August's incident.

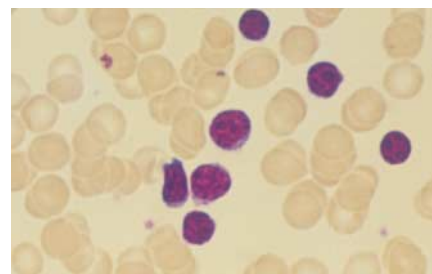
The setback is all the more serious because it arises in a trial that was widely viewed as

gene therapy's only true success. Fisher has so far cured nine boys — including the two who now have leukaemia — out of 11 patients. And when the first child was diagnosed with cancer, some argued that it was an isolated event (see *Nature* 420, 595; 2002).

But Christof von Kalle of the Cincinnati Children's Hospital says analysis of the two boys' cells shows that the same molecular events probably caused the cancers. In both boys, the retroviral vector used to deliver the corrective gene has integrated itself into a stretch of DNA in or near a gene called *LMO2*, which can cause childhood leukaemias. Scientists are now satisfied that the disease behaves enough like leukaemia to describe it as such.

Kalle estimates that the vector is likely to insert itself near to *LMO2* in between 1-in-50,000 and 1-in-100,000 cells. And because each child in the trial receives a dose of about a million corrected cells, each patient could carry at least one cell with the vector near *LMO2*. Kalle is analysing samples from all of the children — about 15 in all — who have been treated in SCID gene-therapy trials to find out whether they are carrying such cells.

In the meantime, an FDA advisory committee on gene therapy will meet on 28 February to decide what the adverse event means for other studies. The American Society of Gene Therapy will also convene a committee



Blood cancer: leukaemia has been detected in two patients who have received gene therapy.

to examine data from human and animal trials of retroviral vectors. And the National Institutes of Health's Recombinant DNA Advisory Committee (RAC) postponed its planned meeting on 17 January, but will convene shortly to discuss these issues.

At their last meetings, both the FDA committee and the RAC recommended that gene therapy in SCID could proceed with appropriate monitoring and informed-consent programmes. RAC chair Ted Friedmann of the University of California, San Diego, says that this approach is "still valid", although it could be revised in light of the new case.

The case is an enormous setback for the field — and for patients with SCID. "The treatment works very well," Fischer says, "but the risk is not acceptable." ■

Meeting aims to find brain's benchmarks for beauty

Jonathan Knight, San Francisco

Are our preferences in music and art learned or innate? Neuroscientists met with artists, musicians and architects at the University of California, Berkeley, on 11 January in an effort to discover some clues to the neurological basis of taste.

Many experts believe that our artistic

choices are entirely dependent on cultural influences. But proponents of neuroaesthetics think that there are pointers to taste, and that studying the brain will help to find them.

The most noted results in this new field have been in music, although even these are inconclusive. Mark Tramo, director of Harvard's Institute for Music and Brain Science, pointed to studies by Marcel Zentner and Jerome Kagan of Harvard University suggesting that infants find more dissonant musical intervals, such as those known as tritones, less pleasing than consonant ones such as perfect fifths (*Nature* 383, 29; 1996).

Tramo said he has shown that perfect fifths produce smoother firing patterns in auditory nerves than tritones do (M. J. Tramo *et al. Ann. NY Acad. Sci.* 930, 92–116; 2001). Although this does not explain why one interval is more pleasing than another, Tramo said, it sheds light on how the auditory system responds to different aspects of music.

Progress is sketchier in the visual arts. John Eberhard, who directs research planning at the American Institute of Architects (AIA), suggested that a set of proportions known as

the 'golden section', which is commonly incorporated in classical architecture, might be universally pleasing to the mind.

But AIA president Gordon Chong pointed out at the meeting that these proportions are completely absent from successful modern buildings such as Frank Gehry's Guggenheim museum in Bilbao, Spain.

Chong remains optimistic that neurobiologists may one day help architects to design soothing hospitals, for example, or stimulating schools. "Many of us look forward to a time when we design spaces more scientifically," he says.

According to Harrison Fraker, dean of Berkeley's College of Environmental Design, current architectural theory holds that taste is learned and relative, not innate. But neuroaesthetes counter that even learned taste must have a neurological basis.

Semir Zeki of University College London, who helped to organize the meeting, says that neuroaesthetics is "taking off in a big way". But for now, he says, "we have much more to learn from artists than they have to learn from us". ■

♦ <http://brain.berkeley.edu/~plaisir/conf.htm>



Strange beauty: Bilbao's Guggenheim museum does not conform to classical proportions.

China set to join international fusion experiment

Washington China has formally asked to join ITER, the US\$5-billion project to build an experimental fusion-energy device, and has offered to contribute about 10% of its costs.

The request was made last week during a visit to China by Robert Aymar, the project's head. A letter confirming the offer has been sent to the project's partners — Japan, Russia and the European Union. "China intends to provide a substantial contribution in kind or in finance to the project," the Chinese science minister Xu Gusnhua said in the letter.

ITER — formerly the International Thermonuclear Experimental Reactor — is a doughnut-shaped machine more than twice the size of existing fusion devices. Since early last year, the United States has been considering rejoining the project, from which it withdrew in 1999, but it has yet to make a formal decision.

Bacteria researcher plagued by fallout from security scare

San Diego An incident that saw 30 vials of plague bacteria declared missing, and sparked a bioterror scare, has ended with a Texan

scientist jailed and accused of lying to the FBI.

Thomas Butler, head of the infectious-disease division at Texas Tech University in Lubbock, and an expert on bubonic plague, reported on 14 January that the vials of plague bacteria, *Yersinia pestis*, were missing. The information sparked a full-scale response from the FBI, but Butler was later arrested for allegedly making false statements to an FBI agent.

After questioning by the FBI, Butler failed a lie-detector test, and admitted that he had accidentally destroyed the vials. Butler, who denies wilfully misleading the FBI, was due to have a bail hearing this week.

Shimadzu sees sales rise after Nobel prize

Tokyo The Nobel Prize may be valued most for its prestige — but it can also prove handy for the bottom line, a Japanese company reports. Shimadzu, based in Kyoto, which makes scientific instruments, says that the success of its researcher Koichi Tanaka in the recent Nobel awards will boost sales of the devices he designed by over 30% this year.

Tanaka's work on mass spectrometers won him a share of the chemistry prize in October last year (see *Nature* **419**, 659; 2002). Shimadzu says it has since had a huge number of enquiries, particularly as the latest

design also went on sale in October. "It was good timing for us," says a Shimadzu official.

The company might even struggle to meet the demand caused by the 'Tanaka effect'. Shimadzu says it has fielded about 140 enquiries, but the UK-based subsidiary that makes the instrument, Kratos Analytical, can only churn out about six of them a month.

Study seeks sea change in approach to oceans

Washington A network of fully protected marine reserves is needed to preserve and revitalize US coastal and marine habitats, a major study recommends.

The report, from the Pew Oceans Commission, says that the combined effects of pollution, overfishing, coastal development

and climate change mean that urgent action is required. The group, made up of researchers, environmentalists, government officials and fishermen, released its findings on 14 January.

If enacted, the proposal would represent a shift in approach towards ocean management, away from traditional, species-specific



Messy problem: US coastal sites need protection.

methods of marine preservation (see *Nature* 418, 718–720; 2002).

• www.pewoceans.org

School space students take ground control

Munich Many schools have pet animals in their classrooms, but German teenagers have an unusually precious clutch of wildlife in their care this week: the Earth-based controls for an experiment on the space shuttle.

Pupils at two schools in Bremen and one in Reutlingen are helping researchers at the German space agency to study the effect of microgravity on a closed aquatic ecosystem. Each school is responsible for a collection of plants, fish, snails and microbes held in a small aquarium. An identical system is currently orbiting 150 miles above on the space shuttle Columbia.

When the shuttle returns on 1 February, the researchers will analyse changes in physiology, morphology and biochemistry of the organisms by comparing them with control measurements taken by the 17- and 18-year-old students. “We found the students very enthusiastic and want to encourage them to choose a research career,” says Reinhard Hilbig, a zoology researcher at the University of Hohenheim in Stuttgart, who is involved in the experiment.

US lab managers face up to congressional scrutiny

Washington A congressional investigation into the management of Los Alamos National Laboratory in New Mexico is to be widened to include two other government labs run by the University of California.

Questions were raised about the running of Lawrence Berkeley National Laboratory and Lawrence Livermore National Laboratory in a letter sent on 10 January by the House Energy and Commerce Committee to the General Accounting Office, which is running the investigation. Both labs are in California and are contracted to the university by the Department of Energy.

Directors at Berkeley and Livermore say they are confident in their labs’ management practices and that they don’t fear the investigation. The initial probe came after the sacking of two independent investigators at Los Alamos, which led to the resignation of the lab’s two top officials earlier this month (see *Nature* 421, 99; 2002).

Nature’s founder gains mark of recognition

London More than 130 years after he founded *Nature*, physicist and astronomer Norman Lockyer has joined the likes of former prime



From left: David Davies (former editor of *Nature*), Philip Campbell (*Nature*’s editor), science minister David Sainsbury and John Maddox (former editor of *Nature*) at English Heritage’s ceremony.

minister Clement Attlee and the poet William Blake in being honoured with a blue plaque from English Heritage. The monument was unveiled on 17 January at the house on Penywern Road in Earls Court, London, where Lockyer lived from 1876 to 1920.

English Heritage, a historical conservation organization, awards blue plaques to “individuals with a strong reputation who have gained recognition through their life and work”.

The plaque was unveiled by British science minister David Sainsbury. “Norman Lockyer was clearly a remarkable man with an incredible breadth of achievement,” Sainsbury said at the ceremony.

Up, up and away

Balloons will soon be able to carry telescopes for months at a time, matching some of the capabilities of satellites, but at a fraction of the cost. Tony Reichhardt charts the course of a new era in ballooning.



Smashing pumpkin: astronomers hope that this balloon design will be able to fly for about 100 days.

This week, near the remote central Australian town of Alice Springs, a group of NASA scientists hopes to unfurl a new kind of balloon. If all goes well, the balloon's clingfilm-thin material will billow outwards as it is pumped full of helium, stretching until it takes on a lobed, pumpkin-like shape. The craft should then drift upwards and begin its journey around the globe, one of the last test flights in a project that aims to transform the way in which astronomers use balloons.

If successful, similar craft could carry telescopes on high-altitude, 100-day missions that would match some of the scientific potential of satellites. The results of the test flight will be keenly watched by astronomers who want to lift their

telescopes above the interfering influence of Earth's lower atmosphere, but who lack the funds for satellite launches. "You can do an awful lot in a 100-day balloon flight," says Giovanni Fazio of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, who has been flying experiments on balloons since the 1950s. "It's like a little satellite experiment at one-tenth of the cost."

The new balloon technology has arrived at just the right time. NASA has tried various schemes to broaden access to space experiments, but with little success. In 1998, it opened competition for University-Class Explorers (UNEX) — low-cost missions that require less than \$15 million for equipment and about \$20 million for launch costs. But the funding was too sparse.

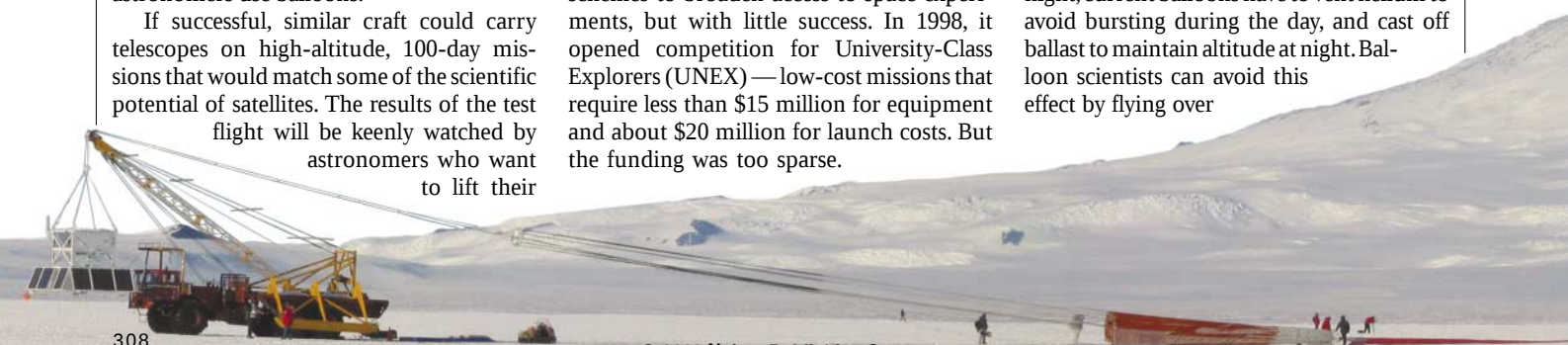
By the time a team has built instruments that can withstand launch and the harsh environment of space, as well as being reliable enough to work for a year in orbit, the \$15 million is spent. NASA envisioned several UNEX missions a year, but the launch two weeks ago of the Cosmic Hot Interstellar Plasma Spectrometer, which will study the hot, diffuse gas that exists between stars, was the first — and could well be the last — UNEX launch.

Compare this with scientific ballooning, where astronomers have for decades lofted instruments into the upper reaches of the atmosphere, close to the boundary with space, for a few days at a time. Most observations taken by this equipment are at wavelengths such as those of X-rays, which are blocked by the atmosphere and so are invisible to ground-based telescopes. Interested parties need only a few hundred thousand dollars, plus another million or so to pay for the crew and other costs associated with the launch. If something goes wrong, it's easy and cheap enough to try again. And, unlike satellites, you get your equipment back at the end.

Positive altitude

Despite the short flight times, NASA's balloon programme has scored some successes. In 1998, the Balloon Observations of Millimetric Extragalactic Radiation and Geophysics experiment — known as BOOMERANG — returned what were then the highest-ever-resolution images of the cosmic microwave background (CMB), the radiation left over from the Big Bang. And all of the devices that flew on the Cosmic Background Explorer (COBE), the pioneering satellite that revealed ripples in the CMB in the early 1990s, had been tested on balloons. COBE was just "three glorified balloon experiments", jokes Stephan Meyer of the University of Chicago, who has worked on both balloon and space-based CMB studies.

Scientific ballooning is now poised to enter the big time. Flying above 99% of the atmosphere, balloons already approach the capability of satellites in terms of generating a clear view for astronomers. Flight times have progressed from less than a week in the 1980s to about a month, and current long-duration balloons can loft about 2,800 kilograms — more than can be carried on all but the most powerful rockets. But because helium gas expands in the heat of day and contracts at night, current balloons have to vent helium to avoid bursting during the day, and cast off ballast to maintain altitude at night. Balloon scientists can avoid this effect by flying over



Alaska or Antarctica in the constant daylight of polar summer. But for trips anywhere else, the mission ends once the ballast is used up.

If the Alice Springs test is successful, all this will change, heralding the era of ultra-long-duration ballooning. The new 120-metre-diameter balloons should fly for about 100 days for a launch cost of about \$2 million. The secret is that, unlike the conventional balloons, the new balloons are sealed to keep the helium at high pressure and their volume constant. Any heat-induced expansion of the helium is absorbed by the balloon's tough skin material and by flexible tendons made from a polymer called polybenzoxazole.

The boost in flight times could be just the first piece of good news. Foreign balloons are currently not allowed to fly over Russian territory, but Vernon Jones, NASA's programme manager for balloons, based at the agency's headquarters in Washington DC, says he has "never been so hopeful" that NASA and the Russian Space Agency will soon reach agreement over letting this happen.

Ways to steer balloons are also being developed, so that in a worst-case failure, a one-tonne telescope payload will not fall out of the sky over populated areas. The odds of this happening are extremely low, says Jones, but some steering capability will be needed to reassure NASA's paymasters that the programme will not spark a disaster.

The simplest idea, proposed by Global Aerospace, a company in Altadena, California, involves using a rudder-like wing to give rough trajectory control. That project has also received interest from NASA's Office of Earth Science, which thinks that such a system would make ultra-long-duration balloons attractive to researchers interested in remote sensing studies. Jack Tueller, an astrophysicist at the Goddard Space Flight Center in Greenbelt, Maryland, and NASA's balloon project scientist, suggests that adding other devices, such as propellers, could also enhance steering capability.

The 100-day flights will open up new possibilities for balloon research, says Fiona Harrison, an astrophysicist at the California Institute of Technology in Pasadena. She is the principal investigator for the balloon-borne High Energy Focusing Telescope (HEFT), which later this year will use state-

of-the-art optics



Vernon Jones: keen to send balloons over Russia.

and detectors to observe supernova remnants and active galactic nuclei during flights a few days long. Over such limited periods, says Harrison, instruments can only make "proof of science" observations, pointing to a few quick targets, and then returning home. But in three or four months, an experiment such as HEFT could study as many as 100 sources and begin to draw more substantial conclusions, says Harrison. And it would carry the latest sensor technology — giving it an advantage over satellites, which usually require instrument design to be fixed early in a mission's development.

Lofty ideas

If the flightworthiness of NASA's new balloon is verified by the Alice Springs launch, the first to benefit will be the University of Maryland's Cosmic Ray Energetics and Mass (CREAM) experiment, scheduled to launch from the National Science Foundation's McMurdo Station in Antarctica this December. CREAM will study fast-moving subatomic particles, called cosmic rays, with energies of 10^{12} – 10^{14} electron volts.

Fazio has longer-term plans. He wants to build the Stratospheric Observatory for Astronomical Research (SOAR), an infrared telescope with a three-metre-diameter mirror — bigger than that on the Hubble Space Telescope. Observing in far-infrared and submillimetre wavelengths for 100 days at a time, and able to be launched again and again, SOAR would conduct sensitive, deep-sky surveys of early galaxies and the clouds of dust in which star formation takes place.

But will such ambitious plans actually be realized? Critics of NASA's attitude to balloons, including Fazio, say that the agency treats the balloon programme like an unwanted stepchild. Recent competitions for NASA's Explorer missions, of which UNEX is the cheapest type, offered long-duration balloons as a platform for the first time. But Fazio and Harrison complain that the dice were loaded. In rating the proposals, NASA considered overall science capability, not 'science per dollar', which put balloons at a disadvantage. Few balloon payloads were proposed, and none was selected.

The competition with satellites may even out once ultra-long flights are demonstrated. If NASA selected a balloon payload for Explorer funding, says Meyer, "it would be an enormously important shot in the arm". Another way to level the playing field would be a balloon-only programme, says Harrison. Either solution, say advocates of balloon studies, would indicate that NASA sees ballooning as a viable high-technology option.

But embracing new technologies could cause problems. Cheapness is ballooning's strongest selling point, although ultra-long-duration missions will cost more than existing ones. Even so, at between \$1 million and \$2 million, these balloons will cost about a tenth of equivalent satellite launches. But the level of technology will have to change, says Meyer. Instruments must be more reliable if they are to operate for 100 days, and researchers will want more advanced devices to take advantage of the longer flights, including telescopes that are even larger than those currently launched by rockets.

Total costs for SOAR, therefore, could reach \$30 million — still cheap by satellite standards, but a far cry from existing balloon experiments. In recognition of this, the decadal review of astronomy by the National Academy of Sciences, published in May 2000, recommended that an extra \$35 million per year be made available for balloon studies.

The scientists developing the concept of ultra-long-duration ballooning are undaunted by potential budgetary pitfalls. "I never expected that building a whole new balloon programme would be easy," says Tueller, who is confident that the payoff will be worth the trouble. If he is right, this month's Australian flight could be the first of many. ■

Tony Reichhardt writes for *Nature* from Washington.

NASA Balloon Program Office

♦ www.wff.nasa.gov/~code820

Ready for take off: many scientific balloons are launched in the Antarctic.





Every day, as a young postdoc, Susan Gasser gazed at a twisted wire skeleton of DNA that was collecting dust in her supervisor's office — a sculptural homage to Jim Watson and Francis Crick's epoch-making paper on the double helix¹.

But Gasser now heads a molecular biology laboratory at the University of Geneva, whose members see DNA in a very different light. By shooting microscopic movies of clumped DNA in the heart of living cells, Gasser and her colleagues have shown the molecule to gyrate like a demonic dancer. For Gasser, the iconic image of DNA as a static double helix is somewhat *passé*: she's now fascinated with the significance of its endless acrobatics.

"The way I think about DNA is different," says Gasser, who is not alone. Watson and Crick transformed biology by revealing the three-dimensional structure of DNA — and in so doing, set the stage for a lasting obsession with its sequence of chemical letters. But 50 years on, researchers are realizing that DNA has a fascinating life in three dimensions — and the fourth dimension of time — that makes it far more than a simple string of code.



Beyond the double helix

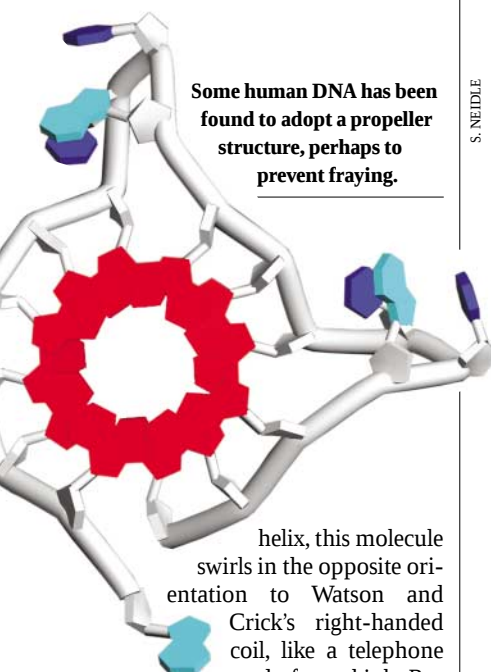
The world of science is gearing up to celebrate the 50th anniversary of Watson and Crick's seminal paper. But there's more to DNA than the pair's iconic structure. Helen Pearson profiles a truly dynamic molecule.

Today's studies paint a more complete picture of DNA by examining the molecule as it coils in the cell's nucleus. In this context, structural biologists now believe that DNA is frequently unfaithful to its famous structure. The double helix, it has emerged, regularly morphs into alternative shapes and weaves itself into knots.

Cell biologists, meanwhile, are exposing the surprisingly dynamic life of the molecule after it crumples up into chromosomes. The latter form fleeting liaisons with proteins, jiggle around impatiently and shoot out exploratory arms. The nucleus, like DNA, was once thought to be fairly static, recalls structural biologist Alexander Rich of the Massachusetts Institute of Technology in Cambridge. "Now we know it to be a very lively place," he says.

Some researchers believe that these mysterious movements may be just as important as the genetic sequence itself in deciding which genes are switched on and off. They even have tantalizing evidence that a failure to coordinate this subcellular waltz could underlie some human diseases. Half a century may have passed since the double helix made its debut, but in some ways, scientists have only just begun to understand this miraculous molecule as it twirls in time and space.

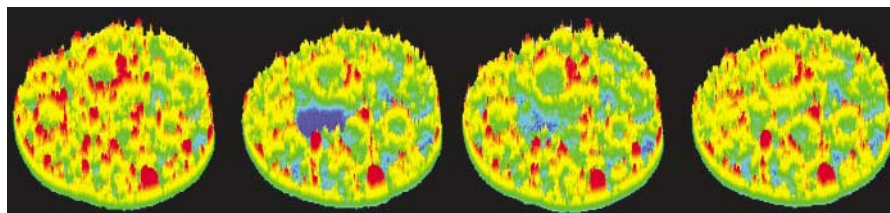
The realization that the double helix isn't the be-all and end-all of genetics stems partly from the discovery by structural biologists of DNA posed in other weird and wonderful shapes. More than two decades ago, Rich first determined the structure of one such variant, called Z-DNA (ref. 2). Although still a double



helix, this molecule swirls in the opposite orientation to Watson and Crick's right-handed coil, like a telephone cord after a kink. But because it was identified in test-tube conditions, the left-handed Z-DNA wasn't considered a significant player in cellular life.

Only recently have researchers found evidence that Z-DNA might be vital in controlling gene activity. In 2001, a team led by Keji Zhao of the National Heart, Lung, and Blood Institute in Bethesda, Maryland, showed that part of the regulatory sequence of an immune-system gene must flip into Z-DNA before the gene can be activated³.

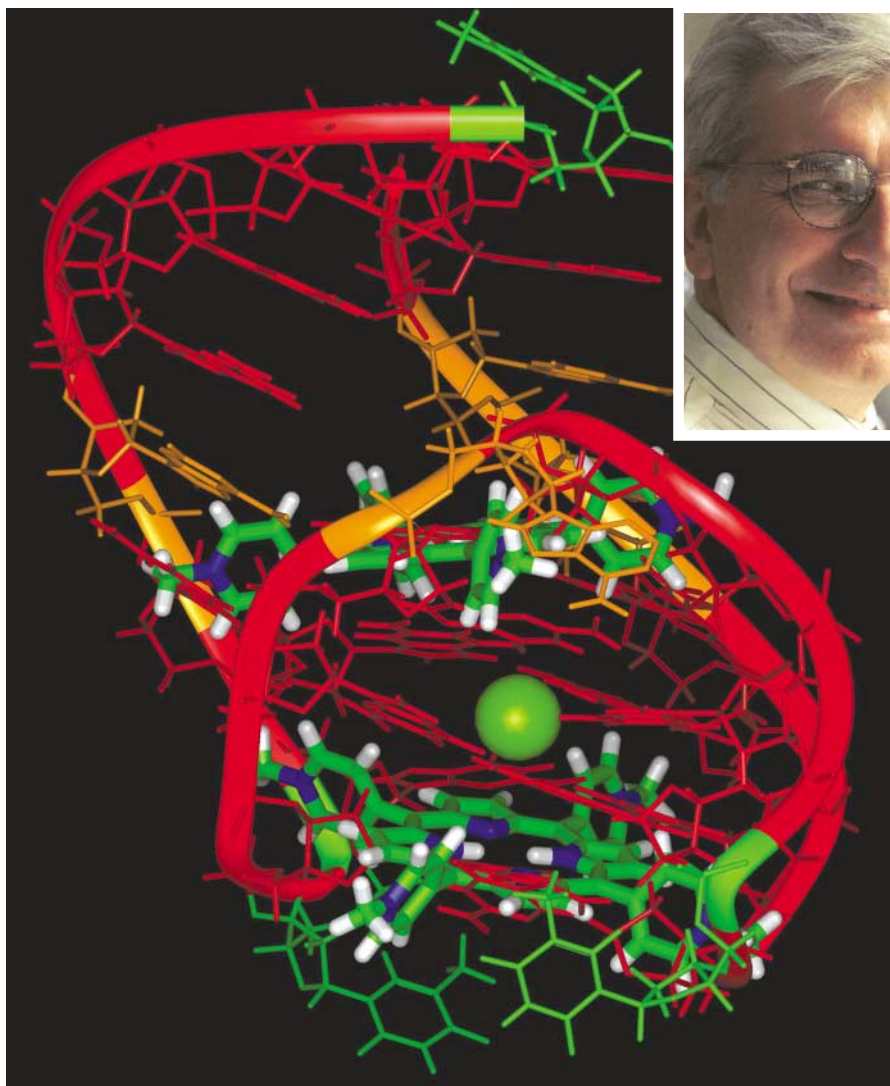
Zhao and other biologists now believe that similar stretches of transiently existing Z-DNA, of which there are perhaps 100,000 in the human genome, may help to switch on genes by making them more accessible to proteins, such as transcription factors, that stimulate gene activity. Building on this idea, Rich



Light show: Tom Misteli (left) has shown the hustle and bustle of proteins in the cell's nucleus. In this time-lapse sequence, a spot (blue) is bleached using a laser, but fluorescent proteins soon diffuse into it.

S. NEIDLE

T. MISTELI/NCI/NIH



Laurence Hurley (inset) thinks this G-quadruplex structure (red) prevents activation of a cancer gene.

already has evidence that the vaccinia virus, which is related to smallpox, may specifically hijack vulnerable Z-DNA, thereby crippling human cells. In unpublished work on mice, he has found that preventing a viral protein from binding to Z-DNA weakens an infection.

Besides the helix's jumps and twists, DNA may also morph into shapes that are a world away from the familiar spiral. Last year, Stephen Neidle of the Institute of Cancer Research in London unveiled one such form that is adopted at the ragged ends of chromosomes, known as telomeres, where the two sister strands give way to a lone string. Using X-ray crystallography, he showed that this single strand can weave itself into a tidy propeller-like loop, which may help to prevent the molecule from fraying away⁴. "It goes well beyond the linear double helix," says Neidle, who is now at the University of London School of Pharmacy.

Neidle's propeller-like structure belongs to a class of departures from helical forms of DNA known as 'G-quadruplexes', which occur in sequences that are rich in guanine, one of the four letters of the genetic code.

Another member of the class may prevent genes from being switched on. Last summer, Laurence Hurley of the University of Arizona in Tucson showed that one type of G-quadruplex forms next to the potent cancer-causing gene *c-MYC*. Negating this structure by mutating its genetic sequence boosted the gene's activity⁵. Hurley suspects that the shape wards off gene-activating proteins, and he is now searching for drugs that could help to stabilize G-quadruplexes and thus serve as anti-cancer agents.

Even in its standard, helical form, DNA is throwing up surprises. The molecule has long been known to form intimate relationships with proteins that help it to fold, and trigger or subdue gene activity. Until recently, these liaisons were thought mostly to be fixed, or to change only slowly with time. But this idea has collapsed, as improved cellular imaging technology has allowed biologists to watch living cells in real time.

The resulting videos exposed an unexpected hubbub in the activity of proteins buzzing around DNA. "It changed the way we thought about the nucleus," says Tom

Misteli of the National Cancer Institute in Bethesda. "The word 'static' is disappearing from our vocabulary."

Misteli's is one of several groups that took the lid off this nuclear ants' nest, using a technique called fluorescence recovery after photobleaching (FRAP). The researchers blasted lasers at live cells containing fluorescent proteins, bleaching out the fluorescence in a small spot of the nucleus. To Misteli's surprise, glowing proteins from elsewhere in the nucleus rushed in to fill the void⁶.

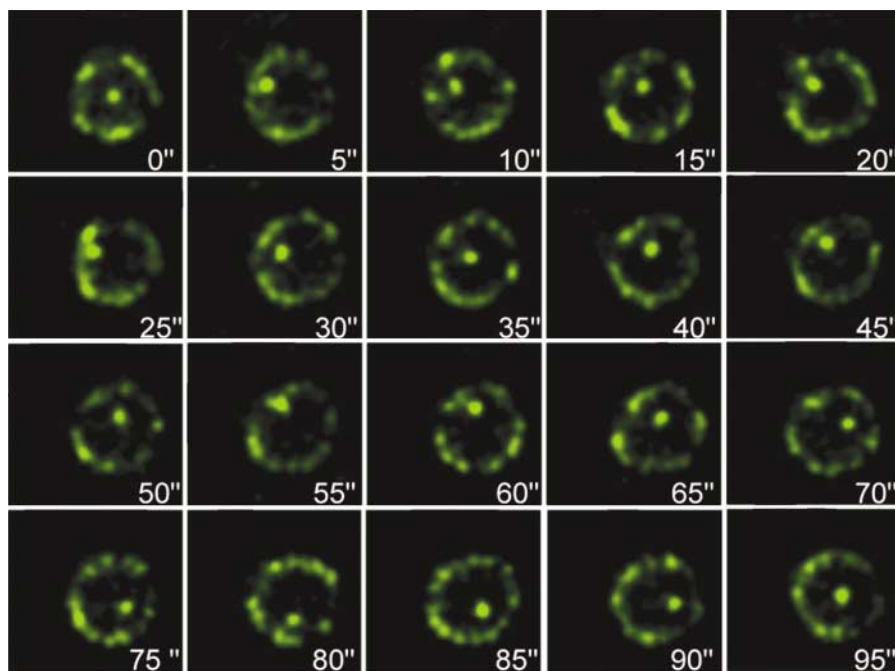
Many researchers now believe that almost all nuclear proteins are scuttling constantly back and forth, moving at speeds that would allow them to traverse the nucleus in as little as five seconds. Even histone H1, a protein that was thought to sit cuddled in the arms of DNA, now appears to attach and detach itself every minute or so^{7,8}. Although one might think that such turmoil would detract from the orderly functioning of DNA, Misteli speculates that this swirl of proteins helps cells to regulate gene activity. It may allow genes to sample constantly from the soup of transcription factors and other regulatory proteins that are milling around the nucleus.

Food for thought

This conjures up a picture of DNA as a tangle of noodles swilling lazily around in a nourishing molecular soup. But in fact, stuffing some two metres of DNA into a nucleus one-millionth of this width requires some exquisitely careful packing. DNA wraps itself around histone proteins to form a nodular structure called chromatin, which in turn coils up like an overtwisted string into globular chromosomes.

Over the past few years, researchers have come to realize that chromosomes, which seem to be carefully arranged in the nucleus, may be positioned so that those that are most important to the cell gain preferential spots. Wendy Bickmore and her colleagues at the UK Medical Research Council's Human Genetics Unit in Edinburgh have shown that human chromosome 18, which carries only a handful of active genes, is relegated to the edge of the nucleus, whereas the gene-packed chromosome 19 sits near the middle⁹. Other researchers have revealed that this organization is preserved in the cells of primates from humans to monkeys¹⁰.

The fact that this chromosomal positioning has been maintained over some 30 million years of divergent primate evolution suggests that the order in which chromosomes are packed in the nucleus is functionally important — but exactly what purpose this packing serves remains unclear. Some scientists suggest that pushing a chromosome to the edge of the nucleus may help to



Room to manoeuvre: these time-lapse images show how a yeast chromosome roams around the nucleus.

hive it off from gene-activating machinery and keep unwanted genes gagged. But Bickmore suspects that peripheral chromosomes may act as defensive 'padding' for the nucleus' valuable genetic cargo, shielding DNA from mutagenic chemicals. "It might be to protect the most important part of the genome from damage," she speculates.

Studies of chromosomal positions used to involve dead cells fixed on microscope slides. But over the past few years, it has also become possible to watch chromosomes in action, thanks largely to a technique for filming a gene's location in living cells that was developed by Andrew Belmont's team at the University of Illinois at Urbana-Champaign. Belmont engineered mammalian cells to carry multiple copies of a bacterial DNA sequence, which can be tagged by a fluorescent bacterial protein¹¹.

Using Belmont's system, researchers have shown chromosomes to be constantly in motion. Chromosomes in cells from yeast to mammals all appear to shimmy around by diffusion within their confined territories, wiggling frantically from second to second¹². Gasser is one of the biologists who has watched these antics, and believes that the shuddering motion helps Misteli's scudding proteins to find their targets in the vast genome. "Before, it was almost inconceivable how a protein would find its binding site," she says.

Chromatin also pulls off some more impressive stunts. Belmont, for example, has engineered mammalian cells to carry a large array of bacterial genes that can be switched on, and tagged, by a fluorescent activating protein. Within several hours of being turned on, the genes unfurl from dense spots of chromatin into diffuse fibres¹³. Applying the same trick to a smaller cluster of bacterial

genes caused its chromatin to move a good distance from the edge of the nucleus towards its interior¹⁴.

Meanwhile, Amanda Fisher and her colleagues at Imperial College, London, have shown that chromatin's wanderings may help to lock unwanted genes in an 'off' configuration. In developing immune cells, Fisher demonstrated that chromatin carrying genes that are no longer needed moved close to regions of inactive DNA called heterochromatin, which is thought to suppress neighbouring genes. If this relocation did not occur, the genes reactivated¹⁵.

Cause and effect?

But Fisher is the first to point out the problem with such observations: they may link DNA movement to changes in gene activity, but they do not prove that one causes another. Some scientists maintain that a wandering genome might simply be a passive consequence of the hustle of proteins around it. But researchers led by John Sedat of the University of California, San Francisco, have provided evidence to the contrary. They studied a mutant version of an eye-colour gene in the fruitfly *Drosophila*, which tends to manoeuvre itself next to genetically mute heterochromatin on the outskirts of the nucleus. The researchers found that, unusually, the dysfunctional gene drags with

it a normal copy of the gene on its partner chromosome — and that, when this happens, the normal copy is also switched off¹⁶.

With such tantalizing signs that chromosomal and chromatin movements may spark or silence gene activity, some scientists are asking whether disruptions in location could trigger disease. Misteli, for example, has gathered evidence that in mouse lymphoma cells, chromosomes 12, 14 and 15 huddle closer together than normal¹⁷. He suspects that their proximity might be what predisposes the cells to become cancerous, by facilitating the abnormal exchange of chromosome regions that can trigger uncontrolled cell division.

Such work implies that patients with a susceptibility to cancer might be diagnosed on the basis of the positions of their chromosomes within the nucleus. With this in mind, Robert Singer of the Albert Einstein College of Medicine in New York has developed a technique that takes a snapshot of the positions of active genes in a single cell. This could be used, for example, to help pathologists to examine a breast biopsy or a suspect skin mole.

The team created 11 fluorescent tags of different colours, and washed them over human cells. Each sticks to the molecules produced by one specific gene, revealing how active it is and its location in the nucleus¹⁸. Looking ahead, others are honing this technique in order to watch human genes in action. They hope to combine multi-coloured tags that will flag not just DNA, but also the messenger RNA molecules churned out by active genes, and the proteins that they encode.

Such developments promise to bring further insights into DNA's enigmatic life in space and time — marking a new chapter in the molecule's history. "Watson and Crick must have thought that sequence was everything," reflects Peter Cook, who studies the structure and function of the cell nucleus at the University of Oxford, UK. "But life is much more complicated than that."

Helen Pearson works in Nature's news syndication team.

1. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738 (1953).
2. Wang, A. H.-J. *et al.* *Nature* **282**, 680–686 (1979).
3. Liu, R. *et al.* *Cell* **106**, 309–318 (2001).
4. Parkinson, G. N., Lee, M. P. H. & Neidle, S. *Nature* **417**, 876–880 (2002).
5. Siddiqui-Jain, A., Grand, C. L., Bearrs, D. J. & Hurley, L. H. *Proc. Natl Acad. Sci. USA* **99**, 11593–11598 (2002).
6. Phair, R. D. & Misteli, T. *Nature* **404**, 604–609 (2000).
7. Lever, M. A., Th'ng, J. P. H., Sun, X. & Hendzel, M. J. *Nature* **408**, 873–876 (2000).
8. Misteli, T., Gunjan, A., Hock, R., Bustin, M. & Brown, D. T. *Nature* **408**, 877–881 (2000).
9. Croft, J. A. *et al.* *J. Cell Biol.* **145**, 1119–1131 (1999).
10. Tanabe, H. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 4424–4429 (2002).
11. Robinett, C. C. *et al.* *J. Cell Biol.* **135**, 1685–1700 (1996).
12. Marshall, W. F. *Curr. Biol.* **12**, R185–R192 (2002).
13. Tumber, T., Sudlow, G. & Belmont, A. S. *J. Cell Biol.* **145**, 1341–1354 (1999).
14. Tumber, T. & Belmont, A. S. *Nature Cell Biol.* **3**, 134–139 (2001).
15. Brown, K. E., Baxter, J., Graf, D., Merksenschlager, M. & Fisher, A. G. *Mol. Cell* **3**, 207–217 (1999).
16. Dernburg, A. F. *et al.* *Cell* **85**, 745–759 (1996).
17. Parada, L. A., McQueen, P. G., Munson, P. J. & Misteli, T. *Curr. Biol.* **12**, 1692–1697 (2002).
18. Levsky, J. M., Shenoy, S. M., Pezo, R. C. & Singer, R. H. *Science* **297**, 836–840 (2002).



A scarlet letter or a red herring?

Genetic discrimination is of little concern compared with existing US healthcare problems.

Sir—A news magazine recently reported on a study published in the *New England Journal of Medicine*¹ showing that a particular combination of common genetic polymorphisms at two different loci increases the risk of congestive heart failure. The writer noted that genetic testing should be helpful in identifying individuals who might benefit from preventive measures, but also lamented that “until we have better safeguards against genetic discrimination by employers and insurers, routine screening is unlikely.”² Blanket assertions of this sort continue to appear in the general media as well as in the peer-reviewed scientific literature³, despite abundant evidence that such discrimination is exceedingly rare, if it occurs at all⁴.

Hall and Rich⁵ and others have amply documented the fact that, for primarily economic reasons, health insurers are not inclined to discriminate on the sole basis of a theoretical risk of future disease. In contrast, an individual who already has a serious illness presents the insurer with an altogether different magnitude of risk, because costly claims are almost inevitable. If an applicant for individually underwritten health insurance is worried about discrimination, it makes more sense to fear a mammogram,

for example, than being tested for a *BRCA* mutation, because tests that detect actual disease are vastly more likely to trigger an adverse underwriting decision. Genetic discrimination in life insurance is also, for sound economic reasons, unlikely to become a significant social problem⁴.

Employers, too, have a financial interest in the current health of present and potential employees. But, here again, examples of genetic discrimination are few. A notable exception seems to be a well-publicized case involving the Burlington Northern Santa Fe Railroad³ in which up to 20 workers who claimed that their jobs gave them carpal tunnel syndrome were unknowingly tested for an obscure genetic marker that the railroad hoped might establish that the problem actually represented a pre-existing (rather than occupational) condition. Although this testing may have been stunningly ill advised, it is remarkable that, when the case is cited as an egregious example of genetic discrimination, it is not mentioned that these workers already had the condition for which the testing was performed.

Genetic discrimination is, according to one definition, “discrimination against an individual or against members of that individual’s family solely because of real or

perceived differences from the ‘normal’ genome of that individual.”⁶ The concept of genetic discrimination becomes meaningless once the definition is broadened to include actions against individuals who already manifest a disease, and yet the fear of such discrimination continues to be validated and perpetuated by the scientific establishment and the media.

Objective examination of the genetic-discrimination issue instead highlights intractable problems in other areas of public policy, notably the need for a universal system of healthcare funding in the United States. Although gene-based medicine will ultimately succeed or fail on its merits, public acceptance of the technology in the near future will hinge on the ability of the medical and research communities to demystify the technology and put the risks and benefits into perspective.

William J. Nowlan

The National Life Insurance Company, 1 National Life Drive, Montpelier, Vermont 05604, USA

1. Small, K. M., Wagoner, L. E., Levin, A. M., Kardia, S. L. R. & Liggett, S. B. *N. Engl. J. Med.* **347**, 1135–1142 (2002).
2. *Newsweek* **74**, 21 October (2002).
3. Rothenberg, K. H., & Terry, S. F. *Science* **297**, 196–197 (2002).
4. Nowlan, W. *Science* **297**, 195–196 (2002).
5. Hall, M. A. & Rich, S. S. *Am. J. Hum. Genet.* **66**, 293–307 (2000).
6. Billings, P. R. et al. *Am. J. Hum. Genet.* **51**, 476–482 (1992).

Travel grants available for genetics congress

Sir—Many geneticists are planning to attend the 19th International Congress of Genetics, to be held in Melbourne on 6–11 July 2003. Because 2003 is the fiftieth anniversary of the publication in *Nature* of the seminal paper by Watson and Crick on the DNA double helix (J. D. Watson and F. H. C. Crick, *Nature* **171**, 737–738; 1953), the congress represents a timely vantage point from which to celebrate the advanced state of the art of genetics, to reflect on half a century of its successes and failures, and to debate its future.

International genetics congresses are significant, inspiring occasions for geneticists from developing countries, providing them with a rare opportunity to meet the top international researchers. Hence, the International Genetics Federation (the official congress sponsor), together with the Australian organizers, has financed some travel bursaries to be awarded on a competitive basis to needy geneticists, especially those from developing nations. Application details can be found on the Congress website at

www.geneticscongress2003.com/index.php.

Anthony J. F. Griffiths

International Genetics Federation, 3529 Biosciences Building, University of British Columbia, 6270 University Boulevard, Vancouver V6T, British Columbia, Canada

Spiritual link is part of traditional knowledge

Sir—In presenting traditional knowledge as a commodity, your News story “Tribes query motives of knowledge databases” (*Nature* **419**, 866; 2002) fails to emphasize the spiritual connection between indigenous peoples and Mother Earth. The use of traditional knowledge must reflect the values that are the foundation of the elders’ practices, especially with regard to medicine. This principle must drive any discussion of how to document such knowledge.

The Centre for Traditional Knowledge understands this and is not asking aboriginal groups to make available their databases as implied in your story. Traditional knowledge should be documented in this way only if the communities themselves choose to do so on their own initiative. Our

organization — not part of the Canadian Museum of Nature, though strongly supported by it through shared facilities and access to scientific expertise — is working only to create a database of expert practitioners and their expertise. We have nearly completed a needs assessment to see how to document their names and expertise in a useful and respectful manner.

Many industries, and governmental and aboriginal organizations, would benefit by identifying the expertise of elders through the use of a database. But only those who know how to work with the spirit world and the medicines will determine what, if anything, can be sold. It is hard to imagine a profit-based venture forming a good relationship with a medicinal plant. Practitioners know that respecting the plant is often essential to the efficacy of the medicine, which is not a miracle chemical compound but a measure of curative energy that draws its medicinal qualities from the relationship between the plant and the people or the person. And you can’t buy a person’s power.

Lynda Kitchikeesic Juden

The Centre for Traditional Knowledge, Canadian Museum of Nature, PO Box 3443, Station D, Ottawa, Ontario K1P 6P4, Canada

Is a scientific boycott ever justified?

Practical guidance is needed to uphold the universality of science.

Colin Blakemore, Richard Dawkins,
Denis Noble and Michael Yudkin

From time to time, groups of scientists propose a boycott of researchers who are citizens of another country, as a political protest against that country's government. It is not widely known that such discrimination is explicitly forbidden by the International Council for Science (ICSU, www.icsu.org) as contrary to the principle of 'universality of science'. (See also *Nature* **417**, 690; 2002 for a statement by the International Human Rights Network of Academies and Scholarly Societies.) ICSU includes nearly 100 national academies of science and research councils, and 26 international scientific unions. We believe that these constituent organizations should do more to make their members aware of the universality of science as a central axiom of scientific conduct. We also suggest that the principle should be taught in graduate training, and accepted as the norm for all scientists.

ICSU's fifth statute describes universality of science as follows: "This principle entails freedom of association and expression, access to information, and freedom of communication and movement in connection with international scientific activities without any discrimination on the basis of such factors as citizenship, religion, creed, political stance, ethnic origin, race, colour, language, age or sex." Although this statement contains some imperfections in wording that we should like to see clarified, its prohibition against discrimination on the grounds of citizenship is clear and unambiguous.

Nevertheless, it is obvious that universality of science cannot be an absolutely inviolable imperative. To take an extreme example, suppose that an international boycott of diplomatic, trade and cultural contacts were declared against a rogue regime as the only way to avoid nuclear war. Would not most scientists agree that a boycott would be acceptable?

Might there be less extreme circumstances in which it would be proper to boycott sci-

The universality of science seeks to prevent the use of scientists as pawns in political activity.

tists, or to discriminate against them in some way? In answering this question, it is essential to consider that science is potentially beneficial to everyone; that the value of a scientific discovery is independent of the discoverer's characteristics; and that the continued ability of scientists to cooperate and transcend boundaries is an important symbol of, and impetus to, the breakdown of political divisions.

To expand on this last point, free communication of information and ideas has historically been significant in the liberalization of autocratic regimes — for example, it was a factor leading to the end of totalitarian regimes in Eastern Europe. Authoritarian governments try to suppress the flow of information and ideas, and to control the participation of their citizens in international activities. The task of scientists in other countries is surely not to exclude their colleagues who live under such regimes from international contacts, but rather to draw them into dialogue.

An objective approach

Scientists can cooperate even when they belong to states in dispute. For instance, scientists from China and Taiwan attend international meetings together and usually interact on the friendliest of terms. Such contacts, which make an important contribution towards reducing hostility, are readily formed by scientists, because science aspires, however imperfectly in practice, to relative freedom from emotional content.

For all these reasons, we conclude that the threshold needed to justify a boycott of scientific colleagues elsewhere must be extremely high, fulfilling the following conditions. (1) The circumstances are exceptional, and the boycott is undertaken only after considered and careful scrutiny by scientists internationally, leading to an explicit judgement that it is worth abandoning the principle of universality of science on this occasion for a particular, overwhelming gain. (2) A boycott is not merely a political gesture but an action that would help to change the unacceptable behaviour of a regime. (3) Revulsion against a regime, and a belief in the necessity for exceptional measures against it, are so nearly universal that a

boycott would be widely respected. (4) The proposed boycott is part of an internationally agreed programme of measures that express collective horror against a regime and are necessary to avert some foreseeable disaster.

Consider three examples to clarify these points. In the first, scientists in one country have asked their colleagues elsewhere to boycott them to put pressure on their government. It may be altruistic for scientists strongly opposed to their government to sacrifice their own interests for a deeply felt cause, but they are not entitled to sacrifice the interests of their compatriots. Moreover, the request of a group of scientists in one country is not decisive.

In the second, Dr X is known to have been personally involved in actions that violate human rights. Is it appropriate to impose a boycott on him? To boycott X in response to his own actions does not contravene the universality of science: its appropriateness depends on the circumstances and on the strength of the evidence against him. Anyway, X's actions do not justify a boycott of his compatriots.

Dr Y, writing from a military address in a country that has used chemical weapons against its own citizens, asks for information or materials for use in studying the spread of infectious disease. Are we justified in refusing Y's request? The development of bacteriological weapons contravenes international protocols. If we have good reason to believe Y might be involved in such development, we are both entitled and obliged to refuse the request. But we are not thereby free to discriminate against Y's compatriot who writes for information about an innocuous topic.

Scientists have the same rights as other citizens to oppose policies of which they disapprove by any legal means. They can also seek, within the law, to persuade colleagues elsewhere to protest against a particular government. What the principle of universality of science seeks to prevent is the use of scientists as pawns in political activity. Although there might be extraordinary circumstances that justify discrimination (including boycotts) against scientists of a given citizenship, careful deliberation and collective judgement would be needed to define them. ■

Colin Blakemore*, Richard Dawkins†, Denis Noble* and Michael Yudkin‡ are at the *Laboratory of Physiology, †University Museum, ‡Department of Biochemistry, University of Oxford, Oxford, UK.

This article grew out of a series of meetings of the four of us, convened and chaired by M.Y. Our full report can be found at www.nature.com/nature/cbboycott_full.



Science can help to shape politics, the banner says: 'castration of science means impotence of Russia'.

Life after the helix

How Jim Watson saw the structure of DNA transform biology.

Watson and DNA: Making a Scientific Revolution

by Victor K. McElheny
Perseus/John Wiley: 2003. 400 pp.
\$27.50/£18.99

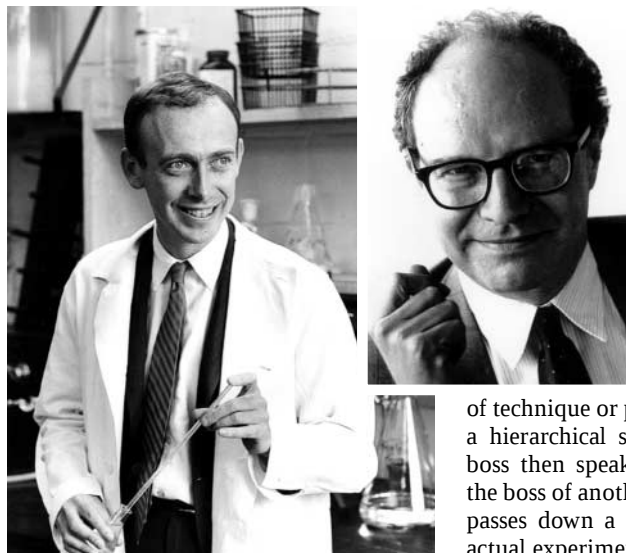
Walter Gilbert

Victor McElheny's biography of Jim Watson captures the growth of molecular biology and the creation of two major institutions that Watson has led: the Cold Spring Harbor Laboratory and the Human Genome Project. It is a vivid portrait of a man who is himself an institution in our science. Covering Watson's childhood in Chicago, the discovery of the structure of DNA with Francis Crick in Cambridge, UK, and his creation of a major laboratory of molecular biology at Harvard, the book contains many insightful comments and quotations. It describes well the development of Cold Spring Harbor from a sleepy, minor research institution to an active centre whose courses, books and meetings have driven the pace of scientific research in modern biology.

The past half-century has truly been the era of DNA, with the publication by Watson and Crick of a model for its structure in 1953, the identification of the genetic code in the 1960s, recombinant DNA and DNA sequencing in the 1970s, biotechnology and the polymerase chain reaction in the 1980s, and the Human Genome Project in the 1990s. The vision that molecules could explain life, and that the most important molecule was that which carries the genetic information from parent to daughter cell, has transformed not just biology but also the world around us. Every organism on Earth is assembled by a set of instructions carried as a linear string of information in a set of DNA molecules, which have an unbroken line of ancestor molecules going back to the first cells that arose upon the Earth.

Watson has been blessed by founding this field, growing the science and shepherding the Human Genome Project in its early days. He is a brilliant, acerbic man, who was amazingly successful not just in his science, but also in leadership and management roles. Although he has a reputation of never suffering fools and likes to shock, he could accommodate himself as needed. McElheny describes how Watson "carefully bends down and unties his shoes and musses up his hair" before going to see wealthy people who might support Cold Spring Harbor.

I met Watson at a party in Cambridge, UK, in the autumn of 1955. We became friends and, in the spring of 1960, when we had both had gone to Harvard, Watson as an



Meeting of minds: Jim Watson and Walter Gilbert (inset) ran a lab together at Harvard in the 1960s.

assistant professor, he invited me to join him and Francois Gros to work on the messenger-RNA experiments. I was a student of theoretical physics, but in the simpler state of science then I could read six papers, follow Watson and Gros around for a day, and then join in the experiments. We sought an unstable intermediate molecule that could carry the information from DNA to the cytoplasmic factories, the ribosomes, which synthesized the proteins. The three of us did the experiments together — one shaking a large flask of bacteria, one holding the stopwatch, and one pouring in 20 mC of radioactive phosphate to label the RNA for 20 seconds before the litres of culture were poured over ice. We worked day and night throughout the summer. We were seeking an elusive target, which many people did not believe existed, and which we would often compare, when speaking to each other, to the search for the neutrino. However, our experiments were ultimately successful (*Nature* **190**, 581; 1961).

For the next decade, Watson and I ran a laboratory together at Harvard. Although he did not have great experimental skills, he had an unusual insight for phrasing important questions and an intuition for the essential new element. We were both driven by a fascination with the new, the flood of discoveries that characterized that period of molecular biology. McElheny's book describes this period vividly through the reminiscences of the students and postdocs involved. When we returned to the laboratory at night after dinner, Watson would look up at the lit windows of our labs standing out against the darkened building, and exult that the

laboratory was active. We would talk to the students and postdocs late into the night.

The students were all set different problems. They were not in competition with each other, but did compete (and collaborate) with other laboratories around the world working on the same problem. In a traditional lab, questions

of technique or principle often rise through a hierarchical structure to the boss. The boss then speaks to his friends, generally the boss of another laboratory, the question passes down a chain of command to the actual experimenter, and the answer, probably distorted by rumour, comes back over the peaks. Watson always tried to cut through this pattern, encouraging his students to make direct contact with the person in the other lab doing the experiment — thus avoiding the hierarchy — to get the answer. He believed that science is done best by independent minds solving problems and interacting directly with other scientists as peers, equals in the passion to understand nature. The students published independently. During the 1960s and early 1970s, Watson did not put his name on papers unless he had personally participated in the experiments.

It was a heroic age, in which the science was done by a few people around the world. Small, informal meetings in broken-down gymnasiums characterized the phage-group or the nucleic-acid Gordon research conferences in the 1960s. Today, each of these fields has expanded: where there were 100 people, today there are up to 4,000. Each experiment that was done by students in the lab in the 1960s has become an entirely separate field of science, and there are now gigantic factories, armies of automatic machines, churning out data that cry out to be understood.

Has our science lost its soul? Has molecular biology, which sought basic answers to how life works in terms of the molecules that made life possible, begun to drown as the new systematics (simply the collection of data) lists all of the different genomes, and now proteomes, solely because they are there? Is the world one of infinite complexity?

One simplifying explanation is evolution: how the individual patterns arose over history. But the contingent nature of evolution means that many of the final patterns in organisms are accidental. However, we must

understand the full details of our particular state if we are to have effective medicine. We see the biological exploration of all the genes and the gene variants of the individual human, the record of the ancestors, migrations and kinships, as relevant to our ability to cure, to prescribe drugs tailored to the individual, or to enhance function.

It seems to me that molecular biology is dead. DNA-based thinking has penetrated the whole of biology, and the separate field no longer exists. Also gone is the attempt to answer broad fundamental questions — how does DNA work? what controls a gene? — by single individuals. However, in fractal fashion, new sciences appear in the details as we continue to learn. Science is both an individual and a collective endeavour. Like the artist, the creative individual finds new discoveries — most often manifest at the moments of breakthrough when an idea reveals a new field of knowledge — that characterize its forefront. Watson was such an individual, and he has stimulated that character in others. ■

Walter Gilbert is in the Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

An astronomical adventure

The Century of Space Science

edited by Johan A. M. Bleeker, Johannes Geiss & Martin C. E. Huber

Kluwer: 2002. 1,868 pp. E675, \$595, £399

Paul Hodge

If a publishing house had mentioned to me that it proposed to publish a compendium of 100 essays that covered all of the important topics in space science, totalling nearly 2,000 pages and written by the world's leading space scientists, I would have expressed enthusiasm but great doubt that it could be done. Space scientists are extremely busy people and it would be almost impossible to find an editor with both the prominence and the patience to see such a mammoth project to completion.

But I would have been quite wrong. Kluwer, the publishers of *The Century Of Space Science*, found not just one but three well-known space scientists to edit the project, and these remarkable men did the job with aplomb. The authors of these well-written and well-illustrated essays include many of the pioneers of space science, as well as many of today's most prestigious practitioners. The result is a truly unique publishing accomplishment: a splendid collection of authoritative reviews that transcends academic disciplines.

Although physically divided into two

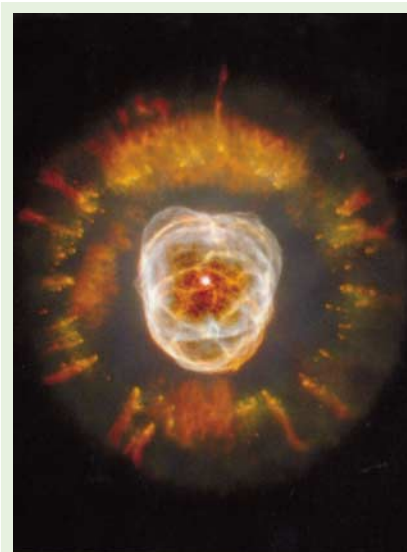
separate, rather heavy volumes, the material is intellectually divided into three sections: the history and development of space science; stellar space physics and cosmology; and Solar System and Earth science. The first section on the history of space science contains so much fascinating material that it would (almost) be worth carrying around in your briefcase to read on the bus or the train, although the volumes are too heavy to qualify as light reading. The later essays also make good reading, with most striking a nice balance between accessibility and technical detail.

The authors include pioneers such as James van Allen, Cornelius de Jager and Donald Hunten, as well as more recent contributors to the literature whose names are more familiar. I will single out the impressive contribution of Arturo Russo, who provides a delightful essay on the prehistory of space science (and also an appendix, which I will mention later). I knew a little about Kepler's interest in space travel, but was unaware of Lucian of Samosota's space-based books *Vera Historia* and *Icaro-Menippus*, which contain detailed descriptions of space travel to the Moon. This essay describes many other historical items that were also new to me.

It is also fun to read William Hartmann's essay on the status of planetary science at the beginning of the space age — how much has changed in the once almost stagnant field of astronomy! Another essay not to be missed is Herbert Friedman's personal account of his involvement with the beginnings of space-based study of the Earth's ionosphere and upper atmosphere.



Launching the space age: rocket pioneer K. E. Tsiolkovsky helped space exploration take off.



Looking stars in the face

Advances in telescope technology have opened up whole new worlds and provided glimpses of distant galaxies. The Hubble Space Telescope, in particular, has provided many stunning images, including this snapshot of the Eskimo Nebula in Gemini — so called because it resembles a face fringed by a fur parka. It actually consists of a pair of elliptically shaped lobes of gas seen end-on. This image is included in *The Times Space: A Photographic Guide to the Universe* by Ian Ridpath (HarperCollins, £16.99).

Space-based results have touched fundamental physics in several ways. Clear, understandable and timely essays about tests of relativity, gravitational lensing, the creation of the elements and the Big Bang are all presented by a list of experts of unarguable distinction. Observational cosmology is also treated well, with comprehensive discussions of such topics as gamma-ray bursts (a hot topic in both senses of the term), quasars and blazars.

One notable landmark is Gustav Tammann's essay on the cosmological constants. Tammann was one of the early advocates of a small Hubble constant (the measure of the rate at which the Universe is expanding). For decades the value of the expansion found by many others seemed to present a contradiction: the expansion age of the Universe seemed to be less than the age obtained by determining the ages of the oldest stars. Tammann's essay presents a clear exposition of this quandary, and shows that current data, based largely on space-based research, provide a reasonable solution, with an age of approximately 14 billion years both for the Universe and also for the stars in it.

A smorgasbord of Earth and planetary

topics occupies the second volume. As an ignoramus when it comes to our most important star, the Sun, I was fascinated by the status of solar physics. Nine separate essays cover virtually everything recently learned about the physics of the Sun, its interior, and how its outer layers and atmosphere interact with its environment.

It might seem as if the planets are not so well treated, as there are only eight essays concerned with planetary research. However, these essays are comprehensive and well designed; they give the reader an interesting tour of the current state of planetary science. There are also three excellent pieces on comets.

In addition to four flavours of index — for acronyms, citations, names and subjects — the back of the second volume includes two appendices that deserve to be widely read, as they bring together information that is otherwise hard to find so neatly in one place. The first is Russo's 'basic chronology of the space age': six pages of space-age highlights, beginning with K. E. Tsiolkovsky's book on space exploration, published in 1903, and ending with the launch of the Soyuz vehicle that took the first crew to the International Space Station in 2000. The second appendix is a tour de force. It is a complete catalogue of space-science launches from 1957 to 2000, with important details of each, listed in a table that occupies an impressive 30 pages of text.

Although the notion of this publishing project would have seemed improbable to me, I have now seen it, lifted it (with some difficulty), and recommend it as an important milestone along the amazing road of space science. ■

Paul Hodge is in the Department of Astronomy, University of Washington, Seattle, Washington 98195, USA.

Ecology out of the blue

Coral Reef Fishes: Diversity and Dynamics in a Complex Ecosystem edited by Peter Sale

Academic Press: 2002. 724 pp. \$99.95, £69.95

Redouan Bshary

Peter Sale's first textbook, *The Ecology of Fishes on Coral Reefs*, became an instant classic for all those who are fascinated by the diversity and complexity of coral-reef fish, their adaptations and their ecology. It left little room for rival textbooks. Now, 11 years on, the steady accumulation of data and ideas has finally warranted a new edition. The title might have changed but fortunately the quality has not.

Coral Reef Fishes is full of fascinating



Evolutionarily quiet: these grunts are little changed from their ancestors 50 million years ago.

details, new evidence and new ideas. For example, we are told that fish on coral reefs 50 million years ago looked very much like they do today (David Bellwood and Peter Wainwright), and that the fish diversity found on coral reefs might largely be due to the constant high temperature, which allows adaptations to exploit low-quality food (Mireille Harmelin-Vivien). Most amazing is the accumulating evidence on the impressive capabilities of reef fish larvae with respect to swimming performance and orientation (Jeffrey Leis and Mark McCormick).

Chapters that describe data at the cutting edge of ecology are well balanced with more 'classical' textbook chapters. There are excellent overviews of reef-fish ecomorphology (Wainwright and Bellwood) and sensory capabilities (Arthur Myrberg and Lee Fuiman), along with critical discussions of methods such as ageing fish using otoliths (Howard Choat and Ross Robertson, Simon Thorrold and Jonathan Hare) and population genetics (Serge Planes).

There is certainly plenty of information on coral-reef fish, but this does have one downside: only a few authors (such as Geoffrey Jones, Julian Caley and Philip Munday) try to fit their data into the general framework of ecology and evolution. More comparisons of reef fish with other taxa could have helped to highlight why studying these fish is not only fascinating but also informative and important.

The final section of the book deals with reef-fish conservation issues and the role that ecology might play. There is a frightening honesty about gaps in our ecological knowledge that should be filled for the effective guidance of management decisions (Peter Sale and Garry Russ). But there is also information on the problem of conserving

sex-changing species (Phillip Levin and Churchill Grimes) and on the live fish trade, which may promote selective hunting of rare species (Yvonne Sadovy and Amanda Vincent).

So what is the future of coral-reef fish ecology? Two major issues emerge from this book. First, a detailed phylogeny down to the species level, including the ages of lineages, will yield the basis for further sophisticated studies of species diversity and the huge variety in adaptations and specializations found in coral-reef fish. Second, increasing our knowledge of the pelagic (open-water) stages of reef fish is of overwhelming importance for several key issues in ecology and conservation. In particular, knowing more about the degree of local retention of larvae has implications for population dynamics, the potential for sustainable local fishing licences and the design of marine protected areas.

Studies of the pelagic stages might also attract the attention of behavioural ecologists. As it stands, coral-reef fish are suitable subjects for behavioural questions only where fitness proxies (such as energy gain and growth rate) are accepted, such as optimal foraging and cooperation. The study of fish larvae opens the possibility of life-history studies and may even allow assessments of individual fitness. But such projects on fish larvae are likely to be labour-intensive and costly. So, to expand Christopher Peterson and Robert Warner's view, what it takes for behavioural ecologists to join the club is a willingness to think big and to get wet. ■

Redouan Bshary is at the Max Planck Institute for Behavioural Physiology, Seewiesen, 82319 Starnberg, Germany, and in the Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK.

Of fossils and fusilli

Nature's insight into what makes scientists tick.

Summarize yourself in the form of a title of a paper in Nature.

Awkward palaeobiologist proved wrong on multiple counts continues to spin out ideas even others find interesting.

Who has been the most important mentor in your career?

Harry Whittington. By the greatest good fortune he was lured from Harvard, so the Burgess Shale campaign ended up in the other Cambridge.

What makes a good scientific mentor?

One who reminds you that patience and an eye like an eagle are no bad things; that the success of others is a matter for rejoicing; and who gives space on the same runway.

Whose graduate student would you most like to have been?

William Buckland, reader in geology at Oxford around 1830. One of the great Georgian eccentrics, he made a determined attempt to eat his way through the animal kingdom.

What scientific event changed your career path?

Looking at Charles Walcott's Burgess Shale papers in the geology department at the University of Bristol — a tip-off from my teacher Crosbie Matthews, another unsung hero.

What underrated discovery changed the science in which you work?

The recognition that, incredibly, organisms and parts thereof, which ought to disappear in less than a day, can fossilize. The windows may be narrow, but the views are superb. Second, that the archetypes and hypothetical common ancestors in the text-books bear about as much resemblance to reality as I do to a martian.

What's the one thing about science that you wish the public understood better?

That it is one of the two great human adventures, that it reveals the world as more extraordinary than we can believe (and yet know), and it opens to us more than ever the choices of doing the right (or wrong) thing. Science is part of our humanity.

Is there a 'tyranny of reductionism' in how scientists are trained today?

Given reductionism works most of the time, nobody is going to forfeit the benefits, until we remember we always end up in a cul-de-sac. How to jump over the wall? Read omnivorously, keep computers at bay, and remember that despite appearances the Universe is a rational place. Puzzles die in cul-de-sacs.

What book has been most influential in your scientific career?

Bob Clark's *Dynamics in Metazoan Evolution*; now rather out of date, but presented with a clarity that combined beautifully the ideas of function and evolution. Very close seconds are books by Steve Vogel, on biomechanics, and James Gordon, on structures. Both write with clarity, humour and verve.

What gives you the most job satisfaction now? What are your major frustrations?

Watching the pieces fall together, yet knowing that in ten years' time most of it will be out of date and forgotten. Frustrations? Journal cuts, ever more stupid bureaucracy, creeping manager-think, and flat-earthers of all types.

What literary character would you employ as a postdoc?

Jim Dixon, hero of *Lucky Jim* by Kingsley Amis. Anarchic surveyor of the absurdities of the academic scene.

What's your favourite conference destination, and why?

Almost anywhere in Italy. Why? Good heavens, don't you know?

What music heads the playlist in your car or lab?

Missa Salve Intemerata by Thomas Tallis.

What was the worst/most memorable comment you ever received from a referee?

"Don't publish — your 'graptolite' is part of a huge arthropod."

What book is currently on your bedside table?

Bruno Bettelheims' *The Uses of Enchantment*; Mary Midgley's *Beast and Man*; Tom Wright's *Matthew for Everyone*; Marion Mill Preminger's *The Sands of Tamanrasset*; Adam Nicolson's *Sea Room*; and about 15 others. Exclusions? No science.

Where and when would you most liked to have lived or worked?

Venice, around 1570.

Assuming the dead can be raised and/or time travel exists, who from the world outside science would you most like to have dinner with?

G. K. Chesterton or — if he's busy chatting to Hilaire Belloc — Fra Paolo Sarpi, friend of Galileo and genius-priest of the Venetian Republic.

What's the best piece of advice you've ever received?

Greenland is quite safe, but orange helicopters definitely are not.



Simon Conway Morris

Simon Conway Morris is a professor in the Department of Earth Sciences at the University of Cambridge, UK. His interests are in evolution, including the Cambrian explosion.

You are on a plane behind two students obviously going to the same conference, who start to talk about your work. What do you do?
Ask for a double gin and tonic, and return to my book.

What would you have become, if not a scientist?
Thirty years ago, a bookseller; today, probably a theologian.

Harry Potter or Lord of the Rings?
One exists, the other is just fantasy.

What single discovery, invention or innovation would most improve your life?
A device that made all cars travel at 4 mph.

Name one extravagance you can now get away with because of your eminence.
Taking time off to dream that one day I might own a painting by Paul Nash or Samuel Palmer.

What music would you have played at your funeral?
John Taverner's *Funeral Ikos*, followed by *Sortie in E flat* by Lefébure-Wély.

How would you like to be remembered?
As someone who slipped through the net of the world.

What's just around the corner?
Something very, very surprising that all of us knew was going to happen.

Cancer summed up

Robert A. Gatenby and Philip K. Maini

Experimental oncology is awash with data. In 2001 alone, over 21,000 articles on characterizing, diagnosing and treating malignancies were published. The powerful techniques of molecular biology have demonstrated innumerable cancer-related alterations in the structure and function of macromolecules that control life and death in mammalian cells. At least 200 genes that may promote or prevent cancer have been identified in the human genome. Remarkably, despite this wealth of information, clinical oncologists and tumour biologists possess virtually no comprehensive theoretical model to serve as a framework for understanding, organizing and applying these data.

Heeding lessons from the physical sciences, one might expect to find oncology aggressively, almost desperately, pursuing quantitative methods to consolidate its vast body of data and integrate the rapidly accumulating new information. In fact, quite the contrary situation exists. Mathematical models are typically denounced as “too simplistic” for complex tumour-related phenomena (ignoring, of course, the fact that similar simplifying assumptions are required in most experimental designs). Articles in cancer journals rarely feature equations. Clinical oncologists and those who are interested in the mathematical modelling of cancer seldom share the same conference platforms.

This attitude begins in medical and graduate schools, where curricula often fail to include theoretical analysis or the application of quantitative methods other than statistics. In fact, medical schools have generally eliminated mathematics from admission prerequisites, resulting in a generation of clinical and research physicians who lack expertise in or regard for biological mathematics.

Those of us who apply quantitative methods to cancer also bear responsibility. Too often we are content with work that is entirely phenomenological — ‘curve-fitting’ data — without developing mechanistic models that provide real insights into critical parameters

that control system dynamics. We also tend to focus on narrow mathematical issues, such as existence and uniqueness theorems, that have no discernible medical application. Most importantly, we have not clearly demonstrated to our biologist colleagues a dominant theme of modern applied mathematics: that simple underlying mechanisms may yield highly complex observable behaviours.

So how might mathematical methods help oncology? Consider the well-known model of colorectal carcinogenesis that was developed by Eric Fearon and Bert Vogelstein. This theory correlates a sequence of genetic and epigenetic events with corresponding changes in tissue morphology as it proceeds serially from normal mucosa, to a small polyp, to a large polyp, and eventually to invasive cancer.

The ‘Vogelgram’ represents a word model, grounded in the linear logic that is typical of this approach. It can, however, form the schematic framework for mechanistic, quantitative models that incorporate more realistic properties of biological systems such as stochasticity and nonlinearity. As the outcomes of such interactions cannot be determined by verbal reasoning alone, they must be computed from general integrative models of carcinogenesis.

These models might, for example, adapt methods from game theory and population biology to frame the ‘Vogelgram’ mathematically as a sequence of competing populations that are subject to random mutations while seeking optimal proliferative strategies in a changing adaptive landscape. The phenotypic expression of each mutation interacts with specified environmental selection factors that confer a proliferative advantage or disadvantage. Such models will generate far less predictable (and more biologically realistic) system behaviour, including multiple possible genetic pathways and timelines in the somatic evolution of invasive cancer.

Critical parameters that emerge from mathematical modelling focus attention on issues that require further theoretical and experimental work. These include explicit identification of environmental factors that control clonal expansion and select the malignant phenotype, as well as how these microenvironmental selection parameters are perturbed by successive waves of increasingly transformed cellular populations.

Our own work addresses some of these questions by showing that system dynamics in early carcinogenesis are dominated by normal tissue-growth promoters and inhibitors. Mutations that alter cellular reception, production or processing of these signals will confer a growth advantage, providing a selection mechanism for the early gene mutations

Mathematical oncology

Clinical oncologists and tumour biologists possess virtually no comprehensive model to serve as a framework for understanding, organizing and applying their data.

depicted in the Vogelgram. As the transformed populations expand, substrate limitations due to diminished blood flow and cellular crowding generate new selection parameters that favour cellular populations that are angiogenic and use rapid but inefficient glycolytic metabolism. Thus, properties that are typical of invasive cancers but are not included in the original model flow readily from quantitative methods. This can be further extended through models of tumour–host interactions showing that environmental perturbations induced by acid excretion from glycolytic metabolism can produce the invasive phenotype.

Existing mathematical models may not be entirely correct. But they do represent the necessary next step beyond simple verbal reasoning and linear intuition. As in physics, understanding the complex, non-linear systems in cancer biology will require ongoing interdisciplinary, interactive research in which mathematical models, informed by extant data and continuously revised by new information, guide experimental design and interpretation.

Fortunately, there are some signs of increasing acceptance of mathematical methods in experimental oncology. Analysis of complex genomic and proteomic data has placed bioinformatics in the biological mainstream. ‘Systems biology’ is applying interesting quantitative methods to the web of transcriptional and protein–protein interaction that follow perturbations in biological systems. Perhaps this presages a time in which mathematical oncology will become integral to the study of cancer. ■

Robert A. Gatenby is in the Departments of Radiology and Applied Mathematics, University of Arizona, Tucson, Arizona 85724-5067, USA.

Philip K. Maini is at the Centre for Mathematical Biology, Mathematical Institute, 24–29 St Giles', Oxford OX1 3LB, UK.

FURTHER READING

Adam, J. A. & Bellomo, N. (eds) *A Survey of Models for Tumour–Immune System Dynamics* (Birkhauser, Boston, 1996).

Keener, J. & Sneyd, J. *Mathematical Physiology* (Springer, Heidelberg, 1998).

Fearon, E. R. & Vogelstein, B. A. *Cell* **61**, 759–767 (1990).

Gatenby, R. A. & Gawlinski, E. T. *Cancer Res.* **56**, 5745–5753 (1996).

Simple maths can model a world of complexity.

Dinosaurs take to the air

Richard O. Prum

Flying birds evolved from a group of bipedal dinosaurs. The latest fossil discoveries from China indicate that the dinosaurian ancestors of birds had four wings — and that these animals glided rather than flapped.

Three questions lie at the heart of the debate about the evolution of birds: the origin of the group itself, the origin of feathers and the origin of flight. Progress in reconstructing the evolutionary history of dinosaurs has established a well-corroborated answer to the first question¹. Birds are a lineage of dinosaurs, and are most closely related to dromaeosaurs and troodontids, both of which belong to a group of bipedal, carnivorous dinosaurs called theropods. Since 1997, new fossils and insights from developmental biology have also supported a coherent solution to the second question². Feathers evolved in theropod dinosaurs before the origin of birds or flight through a series of developmental novelties. Now (page 335 of this issue), Xu *et al.*³ report spectacular 124–128-million-year-old fossils from Liaoning, China, that promise to revolutionize discussion of the last question — how did avian flight evolve?

For more than a century, debate on the origin of bird flight has centred on two different hypotheses⁴. According to arboreal theories, flight arose in tree-dwelling creatures through an intermediate gliding stage, an idea that has been supported by the observation that flight is energetically more efficient at higher speeds (when more lift is generated)⁵. Further, the flight stroke in continuous, level flight is simpler than in take-off from the ground. According to the competing, cursorial theories, flight evolved in ground-living animals via a powered running stage. This view is supported by the evidence that birds evolved from a lineage of terrestrial, bipedal theropods, and that many components of the avian flight apparatus evolved originally in a terrestrial context⁴. Moreover, aerodynamic models⁶ of the flight stroke of *Archaeopteryx*, the earliest bird accepted as such, indicate that its wings could have provided thrust as well as lift, and aided the legs in achieving enough ground speed for a running take-off.

In a colourful and prescient paper of 1915, however, William Beebe⁷ proposed that avian flight evolved through a gliding, four-winged — tetrapteryx — stage with wing feathers on both the arms and the legs. Now Xu and colleagues³ describe a small dromaeosaur, *Microaptor gui*, that sports four wings of fully modern, asymmetrical feathers on its forelimbs and legs, and looks as if it could have glided straight out of the pages of Beebe's note-

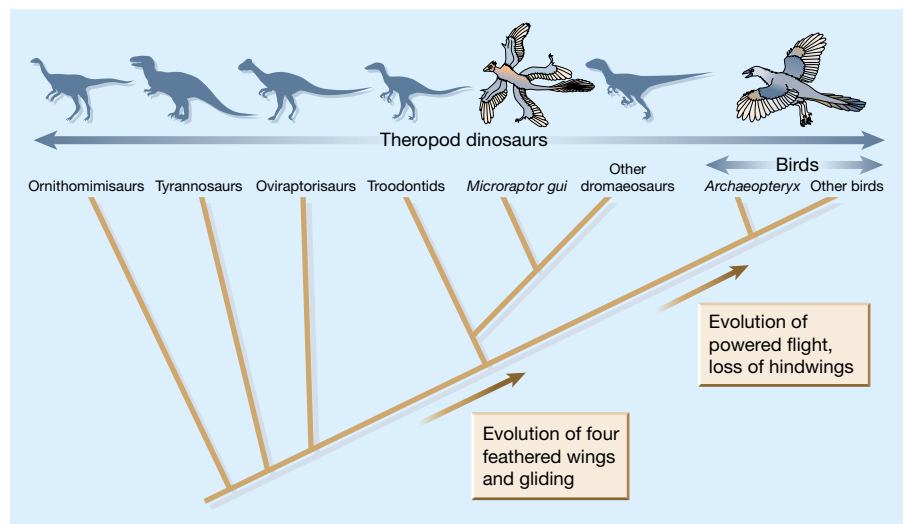


Figure 1 The origin of birds within the theropod dinosaurs¹. Xu *et al.*³ describe a four-winged dromaeosaur, *Microaptor gui*, which they hypothesize glided from tree to tree. They further propose that gliding evolved initially in the four-winged ancestor of the birds, dromaeosaurs, and troodontids. Subsequently, the hindwings were lost with the origin of powered, flapping flight in the ancestor of *Archaeopteryx* and other birds.

books. Although the specimens of *Microaptor* are younger than those of *Archaeopteryx*, *Microaptor* is a basal member — an early evolutionary branch — of the closest relatives of *Archaeopteryx* and other birds (Fig. 1). In support of the arboreal theory, Xu *et al.* propose that the most recent common ancestor of birds and dromaeosaurs was a four-winged creature that lived in and glided among trees.

The evidence of an aerodynamic function for *Microaptor*'s forelimb and leg feathers is excellent. Asymmetrical feather vanes have long been recognized as indicating aerodynamic function in flight or gliding. Furthermore, the feathers of both the forewings and the hindwings increase in asymmetry towards the end of the limb in a striking match with the primary and secondary feathers of a modern bird wing. The discovery of dromaeosaur hindwings was presaged last year by the description of 13-cm-long modern, vaned feathers on the tibia of an even larger, unnamed dromaeosaur from Liaoning⁸, but that specimen lacked enough detail to describe the entire structure. The long tail of *Microaptor* also features a terminal tuft of feathers like that found on other basal dromaeosaurs⁸ and on another theropod, the oviraptoran *Caudipteryx*⁹.

The discovery of a logical functional intermediate provides striking support for

the arboreal–gliding hypothesis of the origin of bird flight. However, substantial questions remain. In particular, how did *Microaptor* actually use its four wings? Perhaps because flapping hindwings are so unlikely, Xu *et al.* conclude that *Microaptor* merely glided, and did not have a powered flight stroke. Palaeontologists and functional morphologists will be eager to study the shoulder and wing anatomy to judge whether *Microaptor* could sustain powered flight. More information is also required on how the animal could have rotated its legs to deploy its hindwings.

Xu *et al.* maintain that all four wings were present in the earliest ancestors of birds and dromaeosaurs, and that with the evolution of powered flight the hindwings were lost in the avian lineage before the advent of *Archaeopteryx*. It is also possible, however, that the hindwings are a unique feature of dromaeosaurs. Palaeontologists will want to re-examine specimens of *Archaeopteryx* for any evidence of a vestigial hindwing (Beebe found none). Regardless of the upshot of those enquiries, there is no doubt that the forewings of *Microaptor*, *Archaeopteryx* and other birds are homologous and had an aerodynamic function.

Xu *et al.* also argue that the extensively feathered legs of *Microaptor* would have been incompatible with life on the ground. The

feathers extend all the way down the leg, much further than they do in Beebe's mythical tetrapteryx. Dragging your wing feathers in the dirt would doubtless be aerodynamically disadvantageous, but it will require detailed reconstructions of *Microraptor's* hindlimbs with feathers attached to rule out the possibility that it could have walked and run.

Finally, although Xu and colleagues report that *Microraptor* has the anatomical features of a dromaeosaur, firm conclusions about the evolution of bird flight will require new systematic analyses incorporating this and other newly discovered theropod species from Liaoning to confirm their phylogenetic position. Sceptics will argue in any case that *Microraptor* and dromaeosaurs are more closely related to modern birds than is *Archaeopteryx* — but then they will also have to address the problem of why a bird that could flap its wings perfectly well would evolve a second pair of wings.

Birds are traditionally considered to be animals with a difference: that is, to be a distinct vertebrate class despite their origins within the reptiles. But advances in palaeontology, phylogenetics and evolutionary biology have erased the anatomical gap between birds and their dinosaur ancestors¹⁰. Now that dromaeosaurs have taken to the air, in the form of *Microraptor*, there remain no major traits that are unique to birds — with

the possible exception of powered flight. Although some may be irked at this lost distinction, the benefits will be a fuller, more integrated understanding of avian biology. This new evidence of an arboreal, gliding stage in the evolution of bird flight complements the evidence of terrestrial evolution in the theropod dinosaurs. Terrestrial theropod dinosaurs had evolved for millions of years before the ancestors of *Microraptor* and the birds took to the trees or to the air. Moving beyond the arboreal versus cursorial debate over the origins of bird flight⁴, the task ahead is to understand which components of the avian flight apparatus evolved in a terrestrial and which in an arboreal context. ■

Richard O. Prum is in the Department of Ecology and Evolutionary Biology, and the Natural History Museum and Biodiversity Research Center, University of Kansas, Lawrence, Kansas 66045, USA. e-mail: prum@ku.edu

1. Sereno, P. *Science* **284**, 2137–2147 (1999).
2. Prum, R. O. & Brush, A. H. Q. *Rev. Biol.* **77**, 261–295 (2002).
3. Xu, X. *et al. Nature* **421**, 335–340 (2003).
4. Padian, K. & Chiappe, L. M. *Biol. Rev.* **73**, 1–42 (1998).
5. Rayner, J. M. V. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 279–287 (Freunde des Jura-Museum, Eichstätt, 1985).
6. Burgers, P. & Chiappe, L. M. *Nature* **399**, 60–62 (1999).
7. Beebe, W. H. *Zoologica* **2**, 38–52 (1915).
8. Norell, M. *et al. Nature* **416**, 36–37 (2002).
9. Ji, Q., Currie, P. J., Norell, M. A. & Ji, S.-A. *Nature* **393**, 753–761 (1998).
10. Padian, K. *Nature* **393**, 729–730 (1998).

foraminiferal abundance indicate that the Holocene was punctuated by a quasi-periodic recurrence of increased monsoon activity on timescales of 1,000 years.

This picture is supported by oxygen isotope analyses² — a part indicator of water temperature — of the same species in a sediment core from the Somali continental margin to the south. The isotope record documents rapid changes in the summer temperature of surface water during the early Holocene, 11,000–6,500 years ago. In turn, both records fit in nicely with isotope analysis³ of a stalagmite, from Hoti Cave in northern Oman, that likewise indicates much variability in monsoon intensity over that interval.

The work of Gupta *et al.*¹ also takes the record of monsoon variability further into the mid- and late Holocene. They detect continued swings in monsoon intensity, including a notable shift from strong to weak activity during the transition from the Medieval Warm Period (AD 800–1300) to the Little Ice Age (AD 1300–1870). This is confirmed by faunal and organic molecular data from the region^{4,5}: monsoon variation, including an increase in strength during the past century or so, is evidently a robust feature of its climatic history.

To add to the excitement, Gupta and colleagues' record shows some resemblance to that seen in the distribution of haematite in Holocene sediments from the North Atlantic. Haematite is believed to be a key indicator of sand debris that was frozen into icebergs and carried across the North Atlantic, and therefore of the rhythmic recurrence of cold spells in the region. A link between cold episodes in the Atlantic and a weakened Asian monsoon has already been documented for the last glacial period, when cold spells — the so-called Dansgaard/Oeschger and Heinrich events — periodically produced arctic conditions in the North Atlantic region^{6–8}. The Dansgaard/Oeschger and Heinrich events caused far more dramatic environmental changes than any of the climatic cycles in the Holocene, so their influence on climatic regimes well beyond the Atlantic region is not surprising. Remarkably, however, the findings of Gupta *et al.* indicate that this linkage continued into and throughout the current warm period of the past 11,000 years, even though the climatic anomalies have been far smaller.

The monsoon system does not operate in isolation, of course. It is only one of many participants in the global dance of climate oscillators and dipoles (see Box 1), and many of these climatic regimes have evidently undergone rapid swings during the Holocene^{9–11}. Following the initial appearance of ice-core palaeoclimate records from Greenland, a belief that climate has been highly stable during the Holocene briefly fluttered through the scientific community.

Global change

Monsoon linkages

Rainer Zahn

An excellent sediment record from the Arabian Sea traces recent patterns in the activity of the Asian monsoon. It reveals both variability in monsoon strength and links with climatic events elsewhere.

The monsoon is the main determinant of environmental conditions over much of Asia, and so affects the most densely populated region on Earth. Differential heating of the north Indian Ocean and the northwest Pacific, and of the Asian land-mass, cause the seasonal reversal of monsoon winds. In summer, these winds blow northwards over the northern Indian Ocean, carrying huge amounts of moisture over the neighbouring land. The ensuing heavy rainfall can have devastating consequences for human life and livelihood. Conversely, agriculture in Asia depends on monsoon rains; and the seasonal upwelling of nutrient-laden subsurface waters, driven by monsoon winds, is essential to the success of coastal fisheries.

Understanding monsoon history and past dynamics is necessary for improving our knowledge of the monsoon system and how it may respond to changing global conditions. On page 354 of this issue¹, Gupta *et al.* take us a step further down that road.

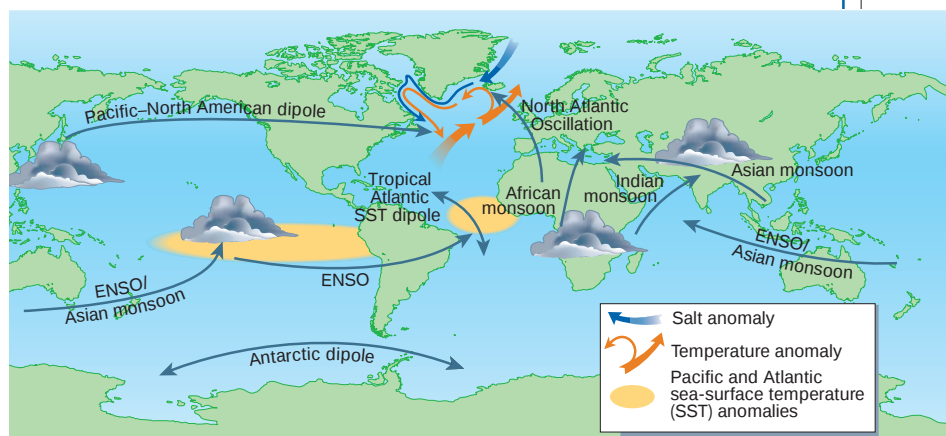
They present a fine-scale palaeoceanographic record, from a marine sediment core from the Arabian Sea, that traces the operation of the Asian monsoon 11,000 years back in time — that is, over the Holocene, the epoch that spans the interval from the end of the last glacial period until the present day.

The record is derived from fluctuations in the abundance of a planktonic organism, a foraminifer, that is known to thrive particularly well in waters that provide an ample supply of food. The link with monsoon intensity is through coastal upwelling, induced by monsoon winds, that stimulates marine biological productivity off Oman, including the growth of this foraminiferal species. Gupta and colleagues' record is particularly valuable because the core comes from a depositional environment in which sediments have accumulated swiftly, so limiting the extent to which burrowing organisms have mixed the sediment column. This record, then, is a very precise one, and the variations in

Box 1 Ocean–atmosphere teleconnections

Climate oscillators and dipoles interact in an intricate dance that redistributes moisture and heat in space and time across the globe. The monsoon subsystems — the Asian, Indian and African monsoons — are connected to variation in sea-surface temperatures in the Pacific. These, in turn, are linked to the El Niño Southern Oscillation (ENSO), so imposing a quasi ENSO-type pattern on monsoon precipitation. The pan-tropical ENSO influences high-latitude atmospheric dipoles, such as the North Atlantic Oscillation and the Antarctic dipole, which respectively determine precipitation patterns in the North Atlantic region and Eurasia, and the extent of sea ice around Antarctica. Together with the Pacific–North American dipole, the North Atlantic Oscillation constitutes the circum-Arctic wave. The Antarctic dipole is the quasi-stationary component of the circum-Antarctic wave.

Some of the heat stored in the tropical oceans, and transported to the poles in surface ocean currents, is released to the atmosphere and creates pressure anomalies. These cause changes in atmospheric convection patterns, and in levels of storminess and precipitation. In the North Atlantic, salinity anomalies are caused sporadically by bursts of fresh water from the Arctic Ocean and the Canadian Arctic seas, in conjunction with swings in the North Atlantic Oscillation. The salt anomalies



stabilize the upper water column and affect oceanic convection patterns, with consequences for the North Atlantic as a source of heat and moisture to the atmosphere. Anomalies in the sea-surface temperature in the North Atlantic are connected with pulses in the transport of warm water in the Gulf Stream, that in turn show linkage with ENSO events in the Pacific. The sea-surface temperature anomalies feed back into the North Atlantic Oscillation, thereby stimulating changes in wind trajectories and precipitation patterns in the wider

North Atlantic region. This then affects the strength of westerly winds into Eurasia, and snow accumulation there, which feeds into Asian monsoon activity.

These teleconnections are evidently highly complicated. We require a better understanding of how they differ in different climatic regimes, and how their stability is affected by changing boundary conditions. One especially large uncertainty is how variations in solar irradiance influence and alter climatic regimes.

R.Z.

That belief is now long dead^{12,13}, and because of both natural and human-induced change we face an uncertain climate future. It will require a concerted effort from those involved in the observational, modelling and palaeoclimate aspects of climate change to integrate knowledge of the various climatic regimes, past and present, to assess the ‘teleconnections’ between them, and the stability of their links through time. Only then will we be able to forecast with any confidence potentially hazardous states of climate regimes^{14,15} — for instance, certain anomalies in sea-surface temperatures in the Pacific and North Atlantic that might promote conditions in which severe flooding occurs in monsoon regions. In providing a more detailed view of monsoon variability in the recent past, and its teleconnections, Gupta and colleagues’ contribution is a good example of how further progress can be made on this subject.

Rainer Zahn is at the *Institució Catalana de Recerca i Estudis Avançats*, and the *Universitat de Barcelona*, *GRC Geociències Marines*, *Departament d'Estratigrafia, Paleontologia i Geociències Marines*, *Campus de Pedralbes, E-08028 Barcelona, Spain*. e-mail: rainer@geo.ub.es

1. Gupta, A. K., Anderson, D. M. & Overpeck, J. T. *Nature* **421**, 354–357 (2003).
2. Jung, S. J. A., Davies, R. G., Ganssen, G. & Kroon, D. *Geochim. Geophys. Geosyst.* **3**(10), 1060; doi:10.1029/2002GC000348 (2002).
3. Neff, U. et al. *Nature* **411**, 290–293 (2001).
4. Anderson, D. M., Overpeck, J. T. & Gupta, A. K. *Science* **297**, 596–599 (2002).
5. Dooze-Rolinski, H., Rogalla, U., Scheeder, G., Lückge, A.

& von Rad, U. *Paleoceanography* **16**, 358–367 (2001).

6. Wang, Y. J. et al. *Science* **294**, 2345–2348 (2001).
7. Schulz, H. et al. *Nature* **393**, 54–57 (1998).
8. Altabet, M. A., Higginson, M. J. & Murray, D. W. *Nature* **415**, 159–162 (2002).
9. Moy, C. M., Seltzer, G. O., Rodbell, D. T. & Anderson, D. M. *Nature* **420**, 162–164 (2002).
10. deMenocal, P., Ortiz, J., Guilderson, T. & Sarnthein, M. *Science* **288**, 2198–2202 (2000).

11. Haug, G. H., Hughen, K. A., Sigman, D. M., Peterson, L. C. & Röhl, U. *Science* **293**, 1304–1308 (2001).
12. deMenocal, P. *Science* **292**, 667–673 (2001).
13. Schulz, M. & Paul, A. in *Climate Development and History of the North Atlantic Realm* (eds Wefer, G., Berger, W., Behre, K.-E. & Jansen, E.) 41–54 (Springer, Heidelberg, 2002).
14. Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. *Science* **295**, 113–117 (2002).
15. Allen, M. R. & Ingram, W. J. *Nature* **419**, 224–232 (2002).

Neurobiology

Fear thou not

Yadin Dudai

Mice lacking a certain neurotransmitter receptor have trouble forgetting scary experiences. This finding uncovers a fear-regulating feedback loop in the brain that might be at work in humans, too.

“Show me a man,” said the Roman writer Seneca, “who is not a slave; one is slave to lust, another to greed, another to ambition, and all men are slaves to fear.” There are few who would seriously challenge this 2,000-year-old conclusion, and fear of fear is probably one of the major motivations for studying emotions in general. Until quite recently, however, it was considered unfashionable for respectable brain scientists to investigate the molecular and cellular bases of the emotions. This has now changed, and we have ample information about the brain circuits that encode fear¹. Writing last month in *Cell*, Shumyatsky

and colleagues² have added important details to the emerging picture. Besides exposing tricks used by the brain to manage fear, the new findings might, in the long run, pave the way to improved treatments for pathological fear responses such as post-traumatic stress disorder.

Study of the cellular and molecular bases of fear requires a model system in laboratory animals. The most popular model is classical (alias pavlovian) ‘fear conditioning’. To understand how this works, it helps to recall how Pavlov trained his dogs in his St Petersburg lab a century ago. In a typical experiment, a dog was presented with a sound, and

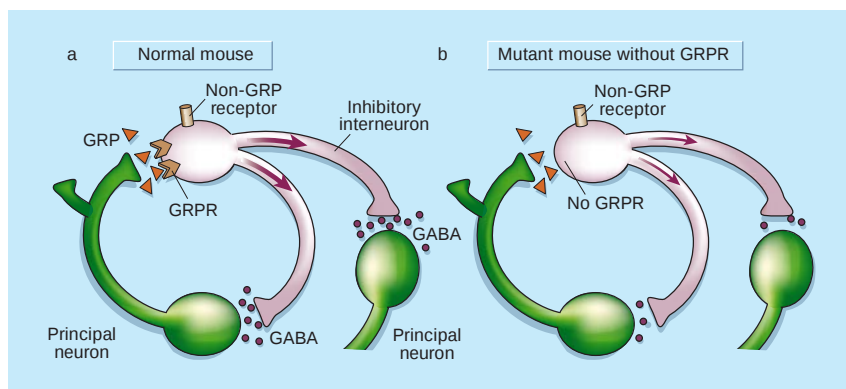


Figure 1 Neuronal circuits and frightened mice. Shumyatsky *et al.*² have proposed the following model for a negative-feedback loop that modulates the activity of the principal neurons in the brain's lateral amygdala, thereby controlling memories of fearful stimuli. **a.** In normal mice, the principal neurons process and transmit information relevant to the formation of fear-associated memories. On activation, these neurons release (among other neurotransmitters) the peptide GRP onto inhibitory interneurons that carry the GRP receptor (GRPR). These interneurons release the inhibitory transmitter GABA onto the principal neurons. Activation of the GRPR stimulates the interneurons to release more GABA, increasing inhibition of the principal neurons. **b.** In mice in which the GRPR is knocked out, the inhibitory interneurons are insensitive to GRP, the feedback loop is severed, and the principal neurons are not properly inhibited. The mice form an abnormally strong and persistent fear memory. The innate level of anxiety and memories of non-fearful events are not affected. (Modified from ref. 2.)

immediately afterwards with meat, which evoked salivation. In the language of psychology, the sound is the conditioned stimulus, the meat the unconditioned stimulus, and salivation in response to meat is the unconditioned response. With time, the dog learned that the sound predicted food, and salivated when presented with the sound alone (the conditioned response). Now, let's substitute the dog with a mouse (or rat), keep the sound, but replace the food with an electric shock to the foot. The result is that the mouse learns to fear the sound. Fear has many manifestations, one of which — freezing of movement — is commonly used as the conditioned response in lab animals.

Part of the neuronal circuitry that subserves fear conditioning is found in the amygdala, a collection of neural structures in the temporal lobe of the brain that controls many aspects of emotional and social behaviour. According to a generally accepted model, information about the conditioned and unconditioned stimuli becomes associated in the lateral nucleus of the amygdala, and output from the amygdala controls the expression of fear (reviewed in ref. 3). Although its exact role is still uncertain⁴, the amygdala is known to be essential for the formation of memory of fearful experiences. Furthermore, in keeping with the conceptual framework that underlies contemporary neurobiology, it is believed that 'fear' in the amygdala to previously innocent stimuli is caused by experience-induced changes in its synapses (the connections between nerve cells).

The molecular basis of fear conditioning is one of the hottest topics in memory research^{5–7}, and Shumyatsky *et al.*² have now tackled certain aspects of it. First, they com-

pared gene expression in different brain areas in mice. This led them to identify a gene coding for a neurotransmitter (a molecule that transmits information between neurons) called gastrin-releasing peptide (GRP), which they found to be preferentially expressed in the principal type of nerve cells in the lateral amygdala and in some interconnected brain areas. The authors then searched for cells that contained the GRP receptor (GRPR) and that could therefore react to GRP. They found the receptor in a subpopulation of amygdalar interneurons — another type of nerve cell — that release the neurotransmitter γ -aminobutyric acid (GABA).

GABA inhibits nerve-cell activity, and Shumyatsky *et al.* found that GRP causes the GRPR-expressing population of interneurons to release more GABA, augmenting their inhibition of the principal cells (Fig. 1a). This effect was abolished in mutant mice in which the GRPR gene was knocked out (Fig. 1b). Moreover, these mutants also displayed enhanced long-term potentiation (LTP) in the neuronal pathway linking the cortex — which perceives and processes fear-associated stimuli, such as the sound in fear conditioning — and the amygdala. LTP is an enhancement of synaptic transmission that persists after repetitive stimulation of a particular neuronal pathway, and it is a major model for learning-related changes in the mammalian brain. Moreover, LTP has been shown to contribute to fear learning *in vivo*⁸. So, this all adds up: an absence of GRPR impairs the inhibition of the principal amygdalar cells by interneurons, and this allows synapses between the neurons of the cortex and amygdala to become more susceptible to LTP.

Together, the data indicate that the lateral amygdala contains a loop that involves the principal neurons and the GRPR-containing inhibitory interneurons. These interneurons regulate the fear-related information that the principal cells process. When the principal neurons fire, they release, among other molecules, GRP, which interacts with GRPR on the interneurons and enhances their inhibition of the principal neurons. So this is a negative-feedback loop.

How is this relevant to behaviour? Shumyatsky *et al.* found that the GRPR-deficient mice displayed a greater and more persistent long-term memory of fear. These mice were not, however, generally anxious, nor did they have a stronger memory of non-fearful events. In other words, the GRP-mediated negative-feedback loop in the amygdala seems to be involved specifically in registering emotionally traumatic experiences.

Several points are noteworthy. First, this work epitomizes the kind of multi-level approach that is essential for real progress in memory research. Second, Shumyatsky *et al.* have dissociated two tiers in a mechanism that regulates the balance between excitation and inhibition in the amygdala. One tier, involving the basal activity of the inhibitory neurotransmitter GABA, controls the innate level of anxiety; the GRPR mutants might have lived happily for ever in a world without painful surprises, because normal basal levels of GABA are released from the inhibitory interneurons in their amygdalas.

The other tier involves the fine-tuning by experience of this balance between excitation and inhibition. This tier constrains the reaction of organisms to emotional trauma. Here GRP is important, and so mutants with a defective GRP-mediated negative-feedback loop overreact to fearful life experiences and have trouble forgetting them, presumably because they lack the mechanism that, in normal mice, quenches fear-induced overactivation of the principal amygdalar neurons. A final point is that only a subpopulation of amygdalar interneurons responds to GRP; other amygdalar circuits, using different neurotransmitters, might fine-tune other emotions.

It will be interesting to see whether these findings translate to humans. Might, for instance, experience-dependent alterations in the GRP-mediated feedback loop contribute to post-traumatic stress disorder following devastating experiences such as rape or a terrorist attack? Might defects in this neuronal system be identified in people who are particularly susceptible to post-traumatic stress disorder? And do some of us overexpress GRP or GRPR, ultimately rendering us less easily frightened than others because we forget scary experiences more quickly? Surely, we are all anxious to know the answers to these questions. But

even before we do, the GRP system is likely to become a focus for research into new, selective, anxiety-relieving drugs. ■

Yadin Dudai is in the Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel. e-mail: yadin.dudai@weizmann.ac.il

1. LeDoux, J. E. *The Emotional Brain: The Mysterious Underpinning of Emotional Life* (Simon & Schuster, New York, 1996).
2. Shumyatsky, G. P. et al. *Cell* **111**, 905–918 (2002).
3. Aggleton, J. P. (ed.) *The Amygdala: A Functional Analysis*

(Oxford Univ. Press, Oxford, 2000).

4. Cahill, L., Weinberger, N. M., Roozendaal, B. & McGaugh, J. L. *Neuron* **23**, 227–228 (1999).
5. Maren, S., Aharonov, G., Stote, D. L. & Fanselow, M. S. *Behav. Neurosci.* **110**, 1365–1374 (1996).
6. Lamprecht, R., Farb, C. R. & LeDoux, J. E. *Neuron* **36**, 727–738 (2002).
7. Ressler, K. J., Paschall, G., Zhou, X. L. & Davis, M. J. *Neurosci.* **22**, 7892–7902 (2002).
8. Rogan, M. T., Staubli, U. V. & LeDoux, J. E. *Nature* **390**, 604–607 (1997).

Thermal physics

Heat in one dimension

Roberto Livi and Stefano Lepri

Heat is transferred along a temperature gradient, from hot to cold, at a rate determined by the thermal conductivity of the material. But is the situation so straightforward in fewer than three dimensions?

Many phenomena in nature occur as the result of some kind of imbalance. For instance, an electric current flows when there is a difference in electric potential along a conductor (such as when an electric field is applied), and heat is transported when there is a temperature gradient between two boundaries of a material. Despite their ubiquity in everyday life, many aspects of such phenomena are still the subject of debate among theoretical physicists. One central issue is the role of spatial constraints, caused by the dimensionality of a system: the response of a system to external forces is intimately related to statistical fluctuations within it, and these, in turn, depend strongly on whether the system is one-, two- or three-dimensional. What happens to energy or charge transport in systems that are effectively one-dimensional, such as a nanowire or a DNA molecule? Onuttom Narayan and Sriram Ramaswamy¹ give their answer in *Physical Review Letters*.

Because of the variety and complexity of specific interactions, simplified microscopic

models are an invaluable tool for the study of transport mechanisms in reduced dimensions (Fig. 1). An example is the old problem of heat conduction. Our own simulations^{2,3} using models of classical low-dimensional fluids and crystals show that the thermal conductivity should increase with the system size. In other words, the larger the system, the more efficiently heat is transported (assuming that the density of the material and the temperature gradient are fixed)—in physical terms, the mean free path of the 'heat carriers' increases with the length of the sample.

Narayan and Ramaswamy¹ are now able to justify this unusual behaviour. They considered a normal fluid, in which the only quantities that vary in time and space are mass, momentum and energy density. They show analytically that the usual concept of thermal conductivity is not well defined for a system of fewer than three dimensions—confirming that space dimensionality is crucial in anomalous transport properties. More specifically, they have estimated that the thermal conductivity diverges as the length of a one-dimensional system increases, following a power law with exponent 1/3; for a two-dimensional system, the divergence is much weaker and logarithmic. But such anomalous behaviour disappears in three dimensions.

Transport anomalies such as this have been found in many microscopic models, including one-dimensional crystals. Although it may seem strange to consider a one-dimensional crystal as a fluid, it is nonetheless well known that mechanical vibrations in a crystal structure can be described by hydrodynamic equations similar to those used in a fluid, as interacting phonons in a crystal behave similarly to particles in a fluid. And as Narayan and Ramaswamy¹ discuss, it is reasonable to expect that the same equations should also be applicable to models of low-dimensional lattices.

For the case of particular lattices of parti-

cles known as Fermi–Pasta–Ulam chains, we have performed extensive computer simulations³ that also indicate a power-law divergence for the thermal conductivity in one-dimensional systems. But our finding for the index of the power law (a value greater than 0.36) is not fully in agreement with Narayan and Ramaswamy's prediction of 1/3. These authors suggest that the discrepancy might arise because, in the one-dimensional case, a fluid and a lattice shouldn't be considered to be exactly equivalent, at least for the intermediate sizes and timescales currently accessible in simulations. In this respect, there are open questions: can the anomalous behaviour actually be described by universal scaling laws, and to what extent does it depend on the nature of the microscopic interactions?

The conceptual challenge is not the only reason for studying energy transport in spatially constrained systems—there is also a variety of real systems in which these anomalies are important. Anisotropic crystals, magnetic chains, polymers and semiconductor films or wires are all examples of systems in which modern experimental techniques can probe the transport properties directly. Single-walled nanotubes (Fig. 1) are known through experiment to have an unusually high thermal conductivity, which is attributed mainly to quasi-one-dimensional lattice vibrations^{4,5}, and it is reasonable to expect that the scaling laws derived for simple models should apply to nanotubes as well⁶. Although, so far, an experimental test is lacking, molecular-dynamics simulations that use realistic energy potentials for the carbon atoms support this idea⁷. If the thermal conductivity did increase with nanotube length in a well-defined way, this would be a very promising feature to use in technological applications, such as the design of components that dissipate heat efficiently in nanocircuits.

In building models of energy-transport processes, the aim is to single out generic physical features, although this might sometimes be at the price of drastic simplifications. Hopefully, in the end the results will go beyond pure academic interest and suggest new ideas for technological applications—perhaps the reader is astonished that so many interesting and innovative ideas are still emerging from classical mechanics. ■

Roberto Livi is in the Dipartimento di Fisica, and INFN and INFN, and Stefano Lepri is in the Dipartimento di Energetica and INFN, Università di Firenze, Florence, Italy. e-mails: livi@fi.infn.it stefano.lepri@unifi.it

1. Narayan, O. & Ramaswamy, S. *Phys. Rev. Lett.* **89**, 200601 (2002).
2. Lepri, S., Livi, R. & Politi, A. *Phys. Rev. Lett.* **78**, 1896–1899 (1997).
3. Lepri, S., Livi, R. & Politi, A. *Phys. Rep.* (in the press); Preprint cond-mat/0112193 (2001), <http://arXiv.org>
4. Hone, J. et al. *Phys. Rev. B* **59**, R2514–R2516 (1999).
5. Kim, P. et al. *Phys. Rev. Lett.* **87**, 215502 (2001).
6. Grassberger, P. & Yang, L. Preprint cond-mat/0204247 (2002); <http://arXiv.org>
7. Maruyama, S. *Physica B* **323**, 193–195 (2002).

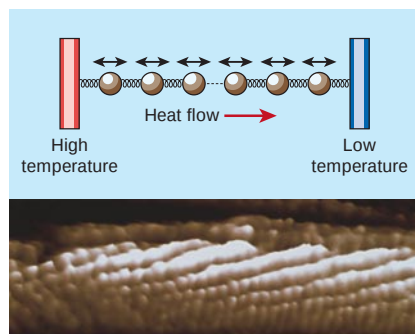


Figure 1 Simple models and real materials. Modelling anomalous heat conduction in one-dimensional systems of interacting particles—such as a chain of oscillating particles that connect two reservoirs at different temperatures (above)—is relevant for understanding heat transport in single-walled carbon nanotubes (below).

A nervous connection

Claude Libert

The molecular details of a connection between the nervous system and the inflammatory response to disease have been uncovered. This suggests new avenues of research into controlling excessive inflammation.

Sepsis is a complex, exaggerated and chaotic version of the usually well-organized inflammatory arm of our immune defences, and kills over 175,000 people each year in the United States alone¹. Although a great deal of time and effort has been spent researching septic shock, it remains difficult to understand and treat. One promising lead was provided two years ago, when it was discovered that there is a connection between inflammation and the involuntary nervous system. The details of this link have, however, been unclear — until now. Writing on page 384 of this issue, Kevin Tracey and colleagues² describe how they identified a receptor protein that is stimulated by the nervous system and which in turn inhibits a key molecular mediator of inflammation and septic shock. This receptor might make a good target for future drugs to treat sepsis.

Inflammation has several roles in the body, one of which is to contribute to the immune system's ability to fight off intruding microorganisms. For instance, molecules that are produced during the inflammatory response increase blood flow to infected areas, or help to recruit immune cells. One way in which inflammation is triggered is in response to lipopolysaccharides — components of the cell walls of many bacteria — which activate the immune system's macrophages. These cells in turn release 'alarm' molecules, namely cytokines, some of which have powerful pro-inflammatory properties. Tumour-necrosis factor (TNF) is one such molecule. This protein can affect nearly all cell types, and has a range of biological activities. For instance, it induces the expression of a large number of genes that encode essential inflammatory molecules (such as other cytokines; enzymes that help to break down the barriers between cells, allowing the migration of immune cells; and adhesion molecules that again enhance immune-cell migration)^{3,4}.

As long as TNF production remains confined to the site of infection, the inflammatory response is clearly beneficial. But once bacteria, and consequently TNF, invade the systemic blood circulation, blood 'poisoning' and sepsis can develop quickly. Furthermore, TNF has been found to be a central mediator of chronic inflammatory disorders such as rheumatoid arthritis and Crohn's disease. So there is much interest in learning how to control the production, release and

activity of TNF. Several means of doing so have been developed (Fig. 1), and have seen some success in treating certain inflammatory disorders⁵. For instance, there are drugs that inhibit the transcription of the TNF-encoding gene into messenger RNA, the translation of the mRNA into protein, or the release of the TNF protein. There are also antibodies and soluble receptors that bind to and block TNF once it has been released. But, although the value of these approaches is beyond doubt, they all take time to work — and time is usually short when treating patients with sepsis.

Tracey's research team has been studying TNF since this protein was discovered (see, for instance, ref. 6). Recently, Tracey's group described another level of control of TNF synthesis — namely by means of the vagus nerve⁷ — thereby providing a new and exciting link between the involuntary nervous system and inflammation. This 'parasympathetic' nerve emanates from the cranium and innervates all major organs in a subcon-

scious way. It is finely branched and is composed of both sensory (input) and motor (output) fibres. This is of relevance because it means that the vagus nerve can on the one hand sense continuing inflammation (presumably by detecting cytokines through receptors on the nerve surface), and on the other hand suppress it. This suppression is efficient and, above all, a good deal faster than the mechanisms mentioned above. Tracey's group found⁷ that, after injecting lipopolysaccharides into rats, electrically stimulating the vagus nerve prevented both the release of TNF from macrophages, and death. Conversely, surgically severing the nerve not only removed this protection but also sensitized the animals to lipopolysaccharide.

But how does the vagus nerve have this effect on macrophages? It was already known that, after this nerve is stimulated, its endings release the neurotransmitter molecule acetylcholine with lightning speed. Macrophages express acetylcholine receptors known as nicotinic receptors, and respond to the released acetylcholine (or the acetylcholine-mimicking nicotine) by suppressing TNF release. But the precise identity of the nicotinic receptors on macrophages was not known. From a therapeutic point of view, this is clearly important to know. It's also very difficult to find out, as the receptors are pentamers containing different combinations of a possible 16 monomers.

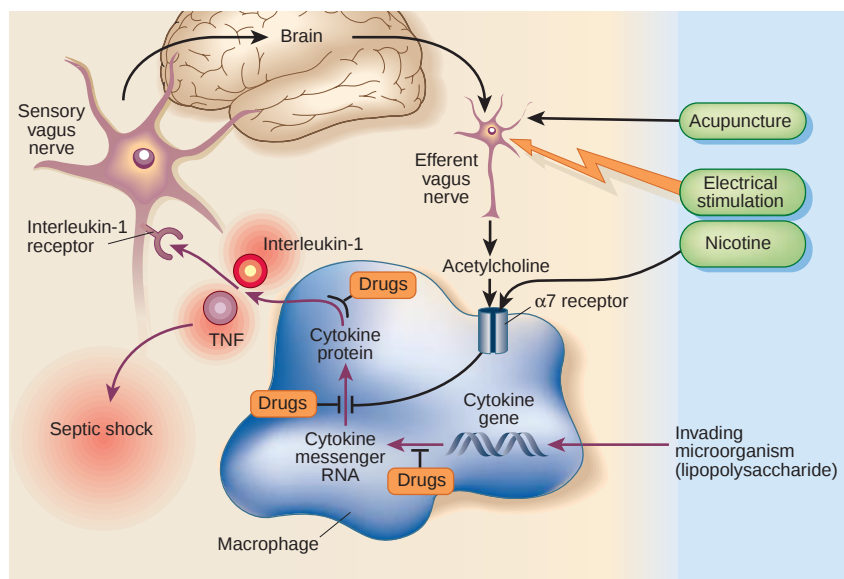


Figure 1 The inflammatory response to microorganisms, and ways of controlling it. Clockwise from lower right: many bacteria contain lipopolysaccharide in their cell walls, which stimulates macrophages. These immune cells then make and release various cytokine ('alarm') molecules, including tumour-necrosis factor (TNF) and interleukin-1. But too much TNF in the blood can be harmful, leading to excessive inflammation and septic shock. Several drugs (orange boxes) inhibit steps in TNF synthesis. In addition, Tracey and colleagues have found that when the vagus nerve detects interleukin-1 (left), it releases acetylcholine (right), which binds to the $\alpha 7$ receptor² on macrophages and inhibits cytokine production. This suggests possible new ways of controlling inflammation: through electrically stimulating the vagus nerve, by acupuncture, or with the use of nicotine (which mimics acetylcholine).

In their latest paper, Tracey and colleagues² pin down the relevant nicotinic acetylcholine receptor: it is one comprising five copies of the monomer $\alpha 7$. They started by using α -bungarotoxin, a molecule that binds to just a subset of receptor monomers, to show that macrophages express the $\alpha 7$ subunit. When the authors blocked the expression of this protein, acetylcholine and nicotine were no longer able to prevent the release of TNF — data that the authors confirmed by studying $\alpha 7$ -deficient mice. In fact, such mutant mice displayed an exaggerated response to lipopolysaccharide in terms of their production of the cytokines TNF, interleukin-1 and interleukin-6. Finally, in a technical *tour de force*, Tracey and colleagues showed that electrically stimulating the vagus nerve of $\alpha 7$ -deficient mice no longer afforded protection against lipopolysaccharide (in contrast to the situation in wild-type mice).

These findings² could have therapeutic implications. The discovery of the connection between the involuntary nervous system and inflammation had already yielded new ideas about treating inflammatory disorders such as sepsis: for instance, a small compound has been developed that can trigger the vagus nerve in rats, thereby reducing inflammation⁸. Looking to the future, it would be interesting to stimulate the vagus nerve electrically in people — as is currently done in thousands of epilepsy patients, showing that the procedure is safe and feasible — and to study the effect on inflammation. More specifically, the new findings suggest that molecules that stimulate the $\alpha 7$ subunit would also be worth developing.

On a different note, nicotine has been found to have powerful immunosuppressive and inflammation-suppressing effects. Of course, the health risks associated with smoking are immense. Yet epidemiological studies indicate that nicotine protects against several inflammatory diseases, such as ulcerative colitis, Parkinson's disease and even Alzheimer's disease. It can also reduce fever and protect against otherwise lethal infection with the influenza virus⁹. The demonstration² that nicotine binds to the $\alpha 7$ subunit on macrophages fleshes out the details of how nicotine produces such effects.

The data also make me reconsider the possibilities and molecular biology of 'alternative' medicine. Pavlovian-type conditioning, hypnosis and meditation are well known (since the beginning of the twentieth century in some cases) to reduce inflammation¹⁰. It might be worth finding out whether these effects, as well as the reported beneficial effects of prayer and acupuncture on inflammation (the last of which is known to depend on acetylcholine)^{11,12}, are mediated by the vagus nerve and the $\alpha 7$ subunit. ■

Claude Libert is in the VIB/RUG Department of

Molecular Biomedical Research, Ledeganckstraat 35, Ghent 9000, Belgium.

e-mail: claudelibert@dmb.rug.ac.be

1. Stone, R. *Science* **64**, 365–367 (1994).
2. Wang, H. *et al.* *Nature* **421**, 384–388 (2003); advance online publication, 22 December 2002 (doi: 10.1038/nature01339).
3. Vassalli, P. *Annu. Rev. Immunol.* **10**, 411–452 (1992).
4. Wielockx, B. *et al.* *Nature Med.* **7**, 1202–1208 (2001).

5. Feldmann, M. *Nature Rev. Immunol.* **2**, 364–371 (2002).
6. Tracey, K. J. *et al.* *Science* **234**, 470–474 (1986).
7. Borovikova, L. V. *Nature* **405**, 458–462 (2000).
8. Bernik, T. R. *et al.* *J. Exp. Med.* **195**, 781–788 (2002).
9. Sopori, M. *Nature Rev. Immunol.* **2**, 372–377 (2002).
10. Talley, N. J. & Spiller, R. *Lancet* **360**, 555–564 (2002).
11. Son, Y. S. *et al.* *Neurosci. Lett.* **319**, 45–48 (2002).
12. King, D. E., Mainous, A. G. III, Steyer, T. E. & Pearson, W. *Int. J. Psychiatry Med.* **31**, 415–425 (2001).

Astronomy

Feeding the first quasars

Laura Ferrarese

Quasars, the oldest known objects in the Universe, are powered by gas falling into black holes at their centres. How black holes formed so early in time has been hard to explain, but a new model might have the answer.

The excitement that has, in recent years, accompanied the study of supermassive black holes mirrors the excitement that followed the discovery of quasars in the early 1960s. Quasars — short for 'quasi-stellar objects' — are characterized by a prodigious outpouring of energy: hundreds of times that of a regular galaxy, but produced in a region that is only a few light days or weeks across. Quasars are also among the most distant objects known to astronomers; as such, the light reaching the Earth from them paints an invaluable picture of the history of our Universe. Supermassive black holes have long been accepted as the only viable energy source for quasars, but only now are we beginning to understand the complex connection between black holes and the formation of galaxies.

Quasars are thought to reside at the centres of massive haloes of dark matter — the mysterious, unseen matter that is needed to explain many features of our Universe. Indeed, this is a crucial assumption that underlies some of the most popular models of black-hole formation. On page 341 of this issue, Barkana and Loeb¹ suggest that if a dark-matter halo is pulling gas towards a quasar at its centre, a distinctive signature should be seen in the light reaching us from the quasar. There are few data to go on so far, but if this signature is confirmed, it would provide the first observational evidence that quasars are embedded in great haloes of dark matter.

In the local Universe, there is now airtight evidence for the presence of supermassive black holes in two galaxies — the Milky Way^{2,3} and the nearby Seyfert II galaxy, NGC 4258 (ref. 4). Compelling, although more indirect, evidence of black holes exists for at least two dozen additional galaxies⁵. These supermassive black holes have masses that range from a few million to a few billion times that of our Sun. Their host galaxies are sufficiently disparate in nature that it is now possible to search for connections between their large-scale properties and the masses of

their central black holes. Some connections have already been identified, most notably between supermassive black-hole masses and the velocity distribution of the surrounding hot stellar component^{6,7}. Another connection may exist between the mass of the central black hole and the mass of the surrounding dark-matter halo⁸.

The demographics of more distant supermassive black holes can be inferred from a census of quasars. In the past few years, the Sloan Digital Sky Survey (SDSS) has dramatically increased the number of known quasars at 'high redshift', where redshift is a measure of an object's recession velocity due to the expansion of the Universe. The SDSS^{9–12} has found several quasars with redshifts larger than 5, including the current record holder at a redshift of 6.43. Assuming that these high-redshift quasars are radiating at the Eddington limit (the maximum luminosity that can be sustained by accretion), their luminosities imply central black holes with masses at least several billion times that of the Sun.

It has been pointed out¹³ that at a redshift of 5 we are looking back in time to when the age of the Universe (about 1 billion years) was approximately equal to the dynamical timescale of a typical galaxy — roughly speaking, the stellar orbital time, or the time it takes a galaxy to communicate with itself through its own gravitational potential. Thus, the very existence of quasars at such high redshifts is a challenge to models of structure formation^{14,15}. Although the details vary, the basic assumption underlying virtually all models is that, at all redshifts, the history of supermassive black holes follows the evolution of galactic dark-matter haloes. In particular, the black-hole mass is assumed to scale with halo mass, and black-hole growth proceeds through gas accretion triggered by galaxy mergers.

A relation between black-hole and dark-halo mass is a feature of models that account for supermassive black holes within this 'hierarchical' framework, in which structure

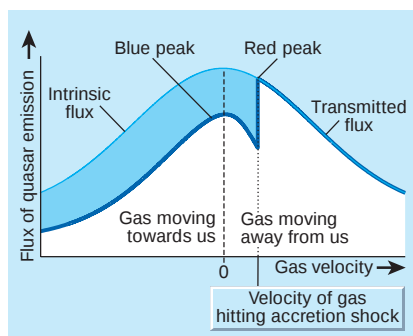


Figure 1 Quasars, black holes and dark matter. Barkana and Loeb¹ show that the observed flux of light from a quasar is modified as gas falling in towards the black hole at its centre, under the gravitational influence of a surrounding halo of dark matter, absorbs some of the radiation. If the same amount of gas is falling in from all directions, the flux as a function of the gas velocity relative to the observer takes on a distinctive 'double-horn' profile. The 'red' peak arises because gas moving away from the observer towards the centre of the quasar hits the accretion shock — a shock barrier due to gas ionization — and inside this barrier the transmitted flux follows the intrinsic spectrum, as the gas no longer absorbs the quasar light. Although Barkana and Loeb have little information to work with so far, their model is in good agreement with the existing spectra of two high-redshift quasars.

formation proceeds from the smaller to the larger galaxies through merging and interactions. Models that successfully reproduce the observed quasar luminosities at high redshifts^{15–17} predict a relation between black-hole and dark-halo mass in remarkable agreement with what is observed in the local Universe.

The observations that triggered Barkana and Loeb's work¹ were spectroscopic studies by the SDSS collaboration of quasars at redshifts of 4.79 and 6.28. The quasar emission lines resulting from excited hydrogen atoms have characteristic 'double-horn' profiles, with distinct red and blue peaks (Fig. 1). Although gas along the line of sight to the quasar could absorb some of the emitted radiation and produce an asymmetric profile, previous attempts at modelling the SDSS quasar spectra in this way have been unsuccessful: absorbing gas that is moving away from the quasar, following the expansion of the Universe, can 'eat away' only the blue side of the emission line profile. Barkana and Loeb realized that if gas were falling in towards the quasar at the centre of the dark-matter halo, the characteristic double-horn profile would be produced.

As gas falls towards the quasar, an increasing volume becomes ionized. Barkana and Loeb's model predicts a sudden change in the fraction of neutral gas around 100 kiloparsecs (roughly 10^{18} km) from the quasar.

At this radius an accretion shock sets in, producing a sharp cut-off in the absorption profile (Fig. 1) — as is seen in the spectra of the SDSS quasars (Fig. 2 on page 342). The velocity of the infalling gas, the radius of the accretion shock and the rate at which mass is being collected towards the centre all depend on the gravitational force exerted by, and hence on the mass of, the dark-matter halo.

Barkana and Loeb's calculations for the two SDSS quasars imply black holes of 10^8 – 10^9 solar masses. Relating these values to the dark haloes, as in hierarchical models, they find that these quasars are embedded in dark-matter haloes each of around 10^{12} solar masses. And, reassuringly, the mass collection rate is large enough that the quasar host galaxies could have been assembled on timescales shorter than the age of the Universe at a redshift of 6.28. But, as Barkana and Loeb admit, more observational data are needed for a rigorous test of their model.

What challenges lie ahead in the study of supermassive black holes? At present, hierarchical models cannot easily account for nuclear black holes with masses smaller than 10^5 – 10^6 solar masses^{16,17}. If such black holes exist, they must be formed through a completely different mechanism — or the entire picture needs to be revised. In fact, the smallest supermassive black hole detected so far is at the centre of the Milky Way (with a mass of 3 million solar masses). Unfortunately, the dynamical signature of lower-mass black holes is subtle at best. Their detection in even the nearest galaxies poses a strong technological challenge as their 'sphere of influence' (the region of space within which the black hole dominates the gravitational potential of the surrounding stars) is smaller than the resolution limits of current instruments. Once again, the ball is in the court of the observers.

Laura Ferrarese is in the Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, New Jersey 08854, USA. e-mail: lff@physics.rutgers.edu

1. Barkana, R. & Loeb, A. *Nature* **421**, 341–343 (2003).
2. Schödel, R. et al. *Nature* **419**, 694–696 (2002).
3. Ghez, A. M., Morris, M., Becklin, E. E., Tanner, A. & Kremenek, T. *Nature* **407**, 349–351 (2000).
4. Miyoshi, M. et al. *Nature* **373**, 127–129 (1995).
5. Kormendy, J. & Gebhardt, K. in *Proc. XX Texas Symp. Relativistic Astrophysics* Vol. 586 (eds Wheeler, J. C. & Martel, H.) 363 (Am. Inst. Phys., New York, 2001).
6. Ferrarese, L. & Merritt, D. *Astrophys. J.* **539**, L9–L12 (2000).
7. Gebhardt, K. et al. *Astrophys. J.* **539**, L13–L16 (2000).
8. Ferrarese, L. *Astrophys. J.* **578**, 90–97 (2002).
9. Fan, X. et al. *Astron. J.* **122**, 2833–2849 (2001).
10. Fan, X. et al. *Astron. J.* **121**, 54–65 (2001).
11. Fan, X. et al. *Astron. J.* (in the press); Preprint astro-ph/0301135, <http://arXiv.org>
12. Schneider, D. P. et al. *Astron. J.* **121**, 1232–1240 (2001).
13. Turner, E. L. *Astron. J.* **101**, 5–17 (1991).
14. Haiman, Z. & Loeb, A. *Astrophys. J.* **552**, 459–463 (2001).
15. Wyithe, S. B. & Loeb, A. *Astrophys. J.* **581**, 886–894 (2002).
16. Haiman, Z. & Loeb, A. *Astrophys. J.* **503**, 505–517 (1998).
17. Haehnelt, M. G., Natarajan, P. & Rees, M. J. *Mon. Not. R. Astron. Soc.* **300**, 817–827 (1998).



100 YEARS AGO

Mr. H. A. Bryden contributes to the *Fortnightly Review* for January an article on the decline and fall of the South African elephant. It appears that the wild elephant has now practically ceased to exist south of the Cunene and Zambesi rivers. About the year 1830, elephant hunting in Cape Colony was prohibited by the British Government. Since that time, remaining herds have been carefully protected, and they still roam the dense jungles of the Knysna Forest and the Addo Bush in large numbers. It is a curious illustration of what a little timely preservation will do for wild creatures that often within a few miles of Port Elizabeth and Mitenhage there are strong troops of these animals, while one may travel elsewhere fifteen hundred miles up country and not succeed in finding a single wild elephant. From *Nature* 22 January 1903.

50 YEARS AGO

The Nutritional Panel of the Food Group of the Society of Chemical Industry has arranged a series of meetings devoted to the subject "Food and the Future"... Mr. F. Le Gros Clark spoke on the "Yardstick of Population" and posed the question, "Will there soon be too many of us?", pointing to an eventual finite limit to the population that could be supported by the world. Famines have always been local and temporary, never of world dimensions, but the real cause for alarm is that no one can foresee in what precise corner of the world hunger may occur... Mr. Le Gros Clark traced the growth in population over the centuries, suggesting an eventual total of 8,000–12,000 million before stability is reached. The problem of their food supply requires not mainly a technological transformation but a revolution in the habits of men and methods of farming — an agrarian revolution on a world-wide scale. The initiation of this revolution, he said, should have started fifty years ago, before the pressure of population had become so great. Nowadays, we cannot be excused because of ignorance, as perhaps could our grandfathers, who had less factual knowledge. Most nations to-day, protested Mr. Le Gros Clark, are not giving the agricultural scientist and food chemist half the chance they should have of meeting the challenge. From *Nature* 24 January 1953.

Developmental biology

Embryos gain foothold in the womb

Science 299, 405–408 (2003)

Many human embryos that fail to develop fall at the first hurdle — when burrowing into their mother's uterus. So the discovery of a molecule that is potentially vital for implantation might help in the diagnosis and treatment of infertility.

Using spare tissue from a private *in vitro* fertilization clinic, Olga D. Genbacev *et al.* found that embryos increase production of a protein called L-selectin on their outer coat of cells six days after fertilization, around the time of implantation. Biopsies showed that at the same time — and indeed every month — the uterine surface displays a set of unique carbohydrates to which L-selectin binds. Polystyrene beads coated in L-selectin stuck to receptive uterine cells.

Genbacev *et al.* suggest that embryos afloat in the uterus may become anchored by the interaction between L-selectin and its targets, triggering the embryos' invasion of the uterine lining. Intriguingly, L-selectin also helps circulating immune cells called leukocytes stick to the side of blood vessels, so they can invade tissue and fight infection.

If this model stands up, measuring the levels of uterine carbohydrates might help to diagnose women who suffer infertility through faulty implantation. Similarly, the efficiency of *in vitro* fertilization might be improved if embryos were screened for a healthy dose of sticky L-selectin.

Helen Pearson

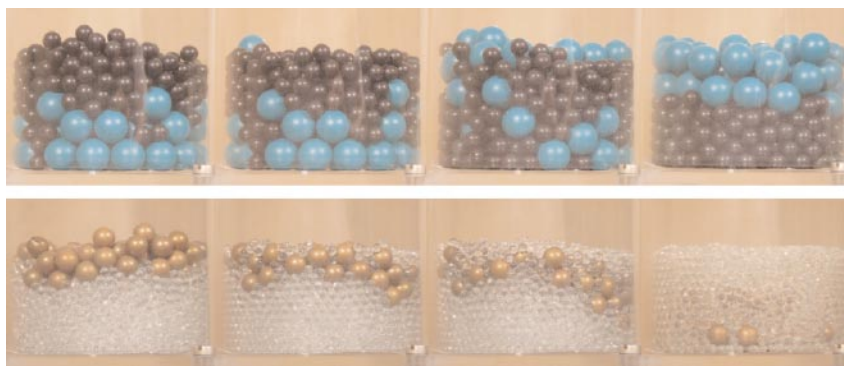
Granular materials

Nut theory gets a shakedown

Phys. Rev. Lett. 90, 014302 (2003)

There are various explanations for the 'brazil-nut effect' — the rising of the largest grains to the top in a shaken granular mixture. These invoke mechanisms such as sieving and convection. But it was pointed out several years ago that the reverse can happen, too: the larger grains may sink.

A theory purporting to explain the crossover between the rising and sinking of the large grains has now been subjected to an experimental test by Andreas Breu and colleagues. They placed various mixtures of two different kinds of spherical bead, made from glass, metal, plastic and wood, in stratified layers and shook them vertically in a clear plastic container (see picture).



Ups and downs: the brazil-nut effect (top panel) and its reverse.

The theory, which claims that the crossover happens when the ratio of particle sizes in a two-component mixture is equal to the inverse ratio of the particle densities, predicts correctly in about 82% of the cases studied.

But the two types of segregation also depended on other factors. A reverse brazil-nut effect could never be produced if the initial lower layer of small grains was too deep. And the same mixture could separate either way, depending on the acceleration of the grains — a function of both the frequency and the amplitude of vibration.

Philip Ball

Geochemistry

Zircons and the early Earth

Earth Planet. Sci. Lett. 204, 333–346 (2002)

Grains of zircon can be dated to more than 4 billion years ago — very early in the Earth's history. The chemical composition of the parent magma from which zircon crystallized is an indicator of past geological conditions, but that parent rock may long since have been destroyed. Can the durable zircons be used to infer the existence, at that early time, of continental crust, and of subduction processes in which crust is dragged down into the mantle below it? Martin J. Whitehouse and Balz S. Kamber have doubts about attempts to do so.

The crucial question is whether the composition of zircon is a reliable reflection of the original magma chemistry. From their analysis of the light rare-earth elements present in 3.8-billion-year-old zircon and its existing parent rock, Whitehouse and Kamber say that it isn't. In a second line of argument, they point out that the light rare-earth-element content of zircon samples from the Moon is akin to that of terrestrial zircons of similar age. There is no evidence that subduction processes have ever occurred on the Moon.

But that won't be the end of the story — zircons are just too precious as a potential

information source about the early Earth for matters to rest there.

Tim Lincoln

Cardiovascular biology

Chilled-out platelets

Cell 112, 87–97 (2003)

Transfusions of blood-clotting cells called platelets can be a life-saver, but there's a problem: platelets have a limited shelf-life because they cannot be stored in the fridge. Karin M. Hoffmeister *et al.* have now discovered why cooling is bad for these cells.

Mysteriously, when chilled platelets are used in transfusions, they are rapidly engulfed and removed by white blood cells. Hoffmeister *et al.* show that a molecule called complement type 3 receptor (CR3) is responsible for the platelets' demise. Mice lacking CR3 on their white blood cells sustain refrigerated platelets for longer. The team suspect that cooling enhances CR3's recognition of a molecule on the platelet surface — a receptor for von Willebrand factor, which is an essential component of the clotting process. They found that chilling platelets makes these receptors cluster together into a more recognizable form.

Hoffmeister suggests that it might be possible to protect refrigerated platelets from white blood cells by blocking the interaction with CR3. Adding enzymes that chemically modify the receptor for von Willebrand factor could do the trick, she says, and allow stored blood to be refrigerated.

The team also speculates that purging chilled platelets is a natural physiological response. Platelets may become 'primed' for activation in the cool outer skin — the location of wounds — and are then removed to avoid potentially harmful clotting throughout the circulation. When whole mice were chilled in the fridge, transfused platelets left their bloodstream faster, the authors found.

Helen Pearson

Frigatebirds ride high on thermals

This bird's bizarre physique and sparse hunting grounds account for its languid lifestyle.

Aspects of the morphology and life history of frigatebirds verge on the extreme, and how they spend their time at sea has been a mystery until now^{1–3}. Here we use data collected by altimeters and satellite transmitters attached to individual frigatebirds to show that these birds are continuously on the wing, day and night — they fly in a succession of climbs and descents, soaring in circles on thermals to heights of up to 2,500 m and gliding down in the direction of travel. The birds' curious morphology and flight patterns result in extremely low costs of foraging, but they also cause them to travel slowly over large distances, putting frigatebirds at an evolutionary extreme that enables them to exploit tropical waters in which prey is scarce.

Frigatebirds (Fig. 1) are designed for an aerial life, having the lowest wing-loading (that is, a relatively large wing area for a low body mass) of any bird and an inability to land on water⁴. Because of their particular morphology, it has been suggested that these sea birds have evolved specific flight strategies to exploit scattered food resources at low cost^{2,3}.

In April 2002, we investigated the vertical and horizontal movements of magnificent frigatebirds (*Fregata magnificens*; Fig. 1) while they were foraging at sea during egg incubation or chick brooding on the nature reserve of Grand Connétable, off the coast of French Guiana in the Atlantic Ocean. To record the birds' altitudes, we fitted seven adults (mass, 1.2–1.5 kg) with continuously recording altimeters (35 g; Suunto Oy; resolution, 3.3 m). We corrected the altimetric data for shifts in atmospheric pressure through recordings taken using a fixed altimeter in the colony.

To record the birds' horizontal move-

ments, we fitted eight individuals with Argos satellite transmitters (Microwave; PTT-100 solar- and battery-powered instruments, 18 and 32 g, respectively) for a total of 42 foraging trips lasting 8.4–93.8 h; loggers and transmitters were taped to the back feathers.

Altimeter readings indicated that frigatebirds foraging at sea flew in a continuous succession of climbs and descents, during both day and night, never remaining at the same altitude for a prolonged period (Fig. 2a). When climbing (average climbing rate, $0.40 \pm 0.14 \text{ m s}^{-1}$; maximum, 3.3 m s^{-1}), birds reached altitudes of up to 2,500 m (average maximum, $304 \pm 34 \text{ m}$); upper altitudes were similar during the day and at night. During descents (average rate, $0.38 \pm 0.15 \text{ m s}^{-1}$; maximum, 2.3 m s^{-1}), birds dropped to an average altitude of $89 \pm 23 \text{ m}$, rarely coming close to the sea surface (on average about once every $7.7 \pm 3.6 \text{ h}$), and then only during the daytime.

The birds' up-and-down flight pattern causes them to move slowly in a horizontal direction. Cross-country ground speeds were particularly low, averaging $10.0 \pm 0.8 \text{ km h}^{-1}$, and even lower at night ($8.6 \pm 2.0 \text{ km h}^{-1}$) than during the day ($11.3 \pm 1.7 \text{ km h}^{-1}$; paired *t*-test, $t=2.5$, d.f.=7, $P=0.036$). Birds covered an average of $223 \pm 208 \text{ km}$ per foraging trip (Fig. 2b), with a maximum range of 27–261 km.

Thermals are generally used by soaring birds by day over land⁵. Over tropical waters in regions affected by trade winds (which blow around the Equator from east to west), convection occurs under cumulus clouds to produce thermals⁶ that allow frigatebirds to soar continuously, day and night. Although the upward air movement in these thermals is slow³, the birds can soar effectively because of their reduced wing



Figure 1 A male magnificent frigatebird (*Fregata magnificens*) in flight off French Guiana. Note the exceptionally large wing area (wingspan, 2.4 m) and inflated gular pouch, which is used as a courtship signal. Frigatebirds spend almost all of their time in flight and are unable to settle on water because of their short legs and feet, and their inadequately waterproofed plumage.

O. CHASTEL

loading, allowing the costs of flight to be greatly reduced.

The constraint on frigatebirds to rely on thermals for flight may explain why these birds are restricted to trade-wind zones, where soaring conditions are optimal throughout the year³. We also tracked two birds during migration outside the breeding season and found that their flight speeds and movements were similar to those during the breeding period, indicating that these probably do not vary over the year.

We have shown that frigatebirds spend the night on the wing — like swifts⁷, the only other species known to do this. Our results also reveal that frigatebirds climb to unexpectedly high altitudes, where the temperature is much lower ($10\text{--}12^\circ\text{C}$ at 2,500 m) than at sea level ($27\text{--}30^\circ\text{C}$). Staying high in the air might be advantageous to frigatebirds, as it may allow them to search for the next thermal, as indicated by the presence of cumulus clouds or of climbing congeners, and to spot feeding opportunities on the sea surface.

Frigatebirds rely mostly on prey that is driven to the sea surface by underwater predators², which are mainly tuna; off French Guiana, estuarine dolphins (tuxuci) are often accompanied by large, multispecies bird flocks⁸ that can also probably be seen by these birds from long distances. Feeding opportunities are rare in these unproductive waters⁹, and our results confirm that frigatebirds only occasionally come close to the sea surface: feeding is presumably infrequent and occurs mainly during the daytime.

The scarcity of resources in tropical waters favours the survival of species whose flight proficiency enables them to travel slowly and economically³ between dispersed concentrations of prey. As a result, frigatebird provi-

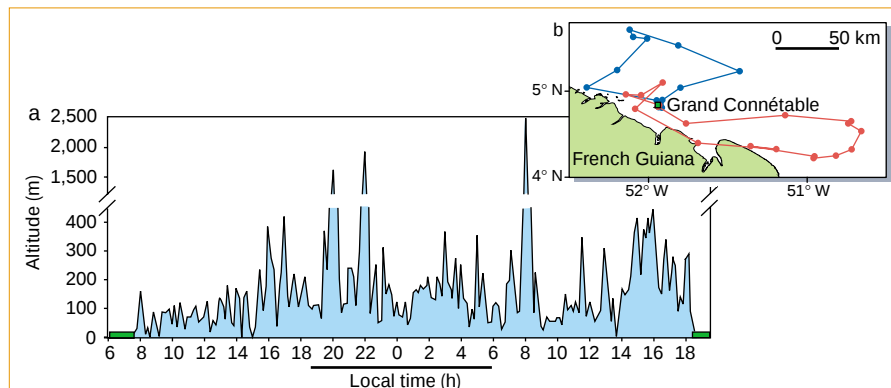


Figure 2 Foraging movements of flying frigatebirds, as revealed by altimetry (for vertical movements) and satellite telemetry (for horizontal movements). **a**, Vertical flight pattern of a sample female during a foraging trip, recorded over two days (24 and 25 April 2002). Black bar shows hours of darkness; green bars represent periods when the bird is in its nest in the colony. **b**, Horizontal movements of two frigatebirds (blue, brooding male; red, incubating female) foraging from the breeding colony (green box) and tracked by satellite telemetry.

sioning is low and their period of parental care is the longest of any bird². Their low reproductive rate and high life expectancy (more than 30 years) make them exceptional even among long-lived birds¹⁰ and, combined with their unusual morphology and foraging strategy, are examples of extreme adaptations to poorly provisioned tropical waters.

Henri Weimerskirch*, Olivier Chastel*, Christophe Barbraud*, Olivier Tostain†‡

*Centre d'Etudes Biologiques de Chizé, UPR 1934

Centre National de la Recherche Scientifique,

79360 Villiers en Bois, France

e-mail: henriw@cebc.cnrs.fr

†Association Arataï, 1 lotissement des Nénuphars,

97357 Rémire-Montjoly, Guyane, France

‡DIREN, 17 place des Palmistes,

97328 Cayenne Cedex, Guyane, France

1. Diamond, A. W. *Condor* **75**, 200–209 (1973).
2. Nelson, B. N. *Living Birds* **14**, 113–155 (1976).
3. Pennycook, C. J. *J. Exp. Biol.* **102**, 307–325 (1983).
4. Orta, J. in *Handbook of the Birds of the World Vol. 1* (eds del Hoyo, J., Elliott, A. & Sargatal, J.) 362–376 (Lynx, Barcelona, 1992).
5. Alerstam, T. *Bird Migration* (Cambridge Univ. Press, Cambridge, 1990).
6. Augstein, E. *Meteorology over the Tropical Oceans* (ed. Shaw, D. B.) (R. Meteorol. Soc., Bracknell, 1980).
7. Lack, D. *Swifts in a Tower* (Methuen, London, 1956).
8. Gould, P. J. in *Pelagic Studies of Seabirds in the Central and Eastern Pacific Ocean* (ed. King, W. B.) 6–52 (Smithsonian Inst., Washington, 1974).
9. Balance, L. T. & Pitman, R. L. *Proc. 22nd Int. Ornithol. Congr.* (eds Adams, N. J. & Slotow, R. H.) 2057–2071 (BirdLife S. Afr., Johannesburg, 1999).
10. Weimerskirch, H. in *Biology of Marine Birds* (eds Schreiber, E. A. & Burger, J.) 115–135 (CRC, Boca Raton, Florida, 2001).

Competing financial interests: declared none.

Pollinator attraction

Crab-spiders manipulate flower signals

Some European species of crab-spider match the colour of the flower on which they lie in wait to ambush insect pollinators, a tactic that is presumed to camouflage them from their intended prey and from predators^{1,2}. Here we show that the coloration of an Australian species of crab-spider, *Thomisus spectabilis*, which is cryptic on the white daisy *Chrysanthemum frutescens* to the human eye, is highly conspicuous to ultraviolet-sensitive insect prey — but that, instead of repelling foraging honeybees (*Apis mellifera*) as might be expected, the contrast of the spider against the petals makes the flowers more attractive. The spider is apparently exploiting the bee's pre-existing preference for flowers with colour patterning.

Visual signals communicated at ultraviolet wavelengths, which are invisible to humans and are therefore more difficult to analyse^{3,4}, may be used by ambush predators to manipulate their prey's behaviour and increase capture success. We have

investigated how *T. spectabilis* interferes with floral signals, and the effect of its visibility on the attractiveness of the flower to pollinating insects.

Under natural light conditions, we presented honeybees with pairs of randomly selected white daisies, one of which carried an anaesthetized spider, and recorded which of the two flowers the bee visited first. We then repeated the experiment using a plastic foil covering on each flower and spider; the cover blocked olfactory cues but was transparent to light of wavelengths greater than 300 nm.

Compared with empty flowers, the presence of white crab-spiders on the petals of daisies evidently attracted honeybees more, in both the presence and absence of olfactory cues (Fig. 1a). This indicates that the bees must have been guided by visual signals alone, and that the visual signal generated by the spider renders the flower more inviting to bees.

To identify this signal, we measured the spectral reflectance from 300 to 700 nm of the flower petals and of the spiders' abdomens. We calculated the colour contrast⁵ of the spiders against the flower petals and

computed the euclidean distances in the colour space of hymenopterans². We found that, compared with the flowers, white spiders reflect a considerable amount of ultraviolet light.

There was also a pronounced difference in the honeybee receptor-excitation values generated by spiders and flowers at ultraviolet wavelengths (Tukey-tests, $P < 0.001$ and $P < 0.001$, respectively), but not in the blue and green regions of the spectrum (ANOVA, $F_{2,74} = 136.8$, $P < 0.001$; Fig. 1b), where receptor excitation is comparable for both (Tukey test, $P = 0.901$). Consequently, instead of being cryptic, as they are to humans, the spiders produce a strong colour contrast that is detectable by their hymenopteran prey (mean euclidean distance in colour space $\pm$ s.e., 0.14 ± 0.01 ; $n = 25$). The values for colour contrast are well above the detection threshold of 0.05 (ref. 2; one-sample t -test, $t_{24} = 7.6$, $P < 0.001$).

We conclude that *T. spectabilis* uses quite the opposite signalling strategy to that known to be used by other crab-spiders^{1,2}. *T. spectabilis* is difficult to perceive from far away, when bees use only their green-receptor signal to detect objects⁶, but is highly conspicuous in the insect visual spectrum when seen at close quarters. Because ultraviolet-reflecting white flowers are extremely rare in nature⁵, the spider will contrast strongly with almost any natural flower.

T. spectabilis will also be just as conspicuous to other flower visitors, as all known pollinating insects, including stingless bees⁷ (which are the spider's most likely Australian native prey), perceive ultraviolet light. We propose that the presence of spiders on flower petals creates a colour pattern that is particularly effective because bees have a pre-existing bias towards it — an idea that is consistent with empirical data showing that bees innately prefer flowers with strongly contrasting markings⁸.

Astrid M. Heiling*†, Marie E. Herberstein†, Lars Chittka‡

*Institute of Zoology, University of Vienna,

1090 Vienna, Austria

e-mail: astrid.heiling@univie.ac.at

†Department of Biological Sciences, Macquarie University, North Ryde, New South Wales 2109, Australia

‡School of Biological Sciences, Queen Mary College, University of London, London E1 4NS, UK

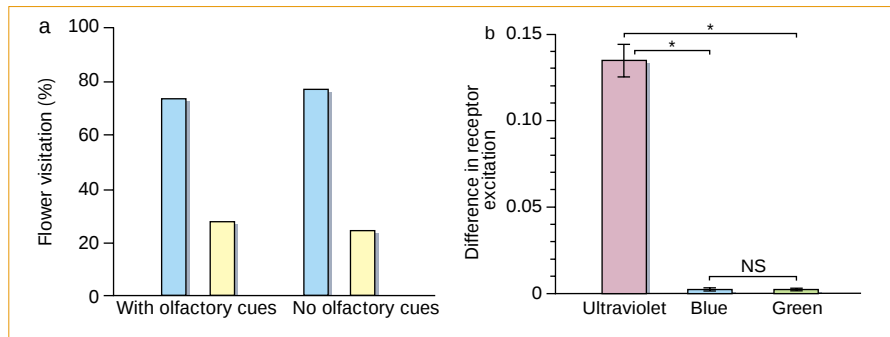


Figure 1 The effect of the presence of the crab-spider *Thomisus spectabilis* on the white daisy (*Chrysanthemum frutescens*) on flower visitation by honeybees (*Apis mellifera*). **a**, Proportion of bees that visit vacant daisies (yellow bars) and daisies occupied by spiders (blue bars) in the presence (binomial test, $n = 33$, $P = 0.0045$) and absence ($n = 25$, $P = 0.0053$) of olfactory cues. All spiders, flowers and bees were used only once. **b**, Difference in honeybee colour-receptor excitation values (mean $\pm$ s.e.; for methods, see ref. 2) between spiders' abdomens and daisy petals at different wavelengths (ultraviolet receptors, $\lambda_{\text{max}} = 345$ nm; blue receptors, $\lambda_{\text{max}} = 440$ nm; green receptors, $\lambda_{\text{max}} = 535$ nm; ref. 7). Tukey test, asterisk denotes $P < 0.001$; NS, not significant.

1. Chittka, L. *Entomol. Genet.* **25**, 181–187 (2001).

2. Thery, M. & Casas, J. *Nature* **415**, 133 (2002).

3. Endler, J. A. *Biol. J. Linn. Soc.* **41**, 315–352 (1990).

4. Bennett, A. T. D., Cuthill, I. C., Partridge, J. C. & Lunau, K. *Proc. Natl Acad. Sci. USA* **94**, 8618–8621 (1997).

5. Chittka, L., Shmida, A., Troje, N. & Menzel, R. *Vision Res.* **34**, 1489–1508 (1994).

6. Giurfa, M., Vorobyev, M., Brandt, R., Posner, B. & Menzel, R. *J. Comp. Physiol. A* **180**, 235–243 (1997).

7. Briscoe, A. D. & Chittka, L. *Annu. Rev. Entomol.* **46**, 471–510 (2001).

8. Lunau, K., Wacht, S. & Chittka, L. *J. Comp. Physiol. A* **178**, 477–489 (1996).

Competing financial interests: declared none.

Four-winged dinosaurs from China

Xing Xu*, Zhonghe Zhou*, Xiaolin Wang*, Xuwen Kuang†, Fucheng Zhang* & Xiangke Du‡

* Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, P.O. Box 643, Beijing 100044, China

† Tianjin Museum of Natural History, Tianjin 300074, China

‡ Radiological Department, People's Hospital, Beijing University, Beijing 100044, China

Although the dinosaurian hypothesis of bird origins is widely accepted, debate remains about how the ancestor of birds first learned to fly. Here we provide new evidence suggesting that basal dromaeosaurid dinosaurs were four-winged animals and probably could glide, representing an intermediate stage towards the active, flapping-flight stage. The new discovery conforms to the predictions of early hypotheses that protoavians passed through a tetrapteryx stage.

For the past few decades, the theropod hypothesis of bird origin has been strongly corroborated by fossil evidence^{1–9} and systematic work^{10–15}. Dromaeosaurids, one of the most distinctive theropod groups, have attracted particular attention since the description of *Deinonychus*¹, owing to its pivotal role in supporting a theropod origin of birds. They, either by themselves or with troodontids^{7,10,11,13–16}, have been regarded as most closely related to birds. A better understanding of these animals is therefore crucial in reconstructing the evolutionary transition towards birds. Recent discoveries from the Jehol Group of western Liaoning, China, suggest that basal dromaeosaurs are small, feathered animals with forelimbs similar to those of *Archaeopteryx*, and feet showing features comparable to those of arboreal birds^{7,8,17}. In 2001 and 2002, we obtained six new basal dromaeosaurid specimens from the Lower Cretaceous Jehol Group at a few localities in Chaoyang Basin, western Liaoning, China. IVPP (The Institute of Vertebrate Paleontology and Paleoanthropology) V13352 and V13320 were identified as a new species of *Microraptor*, TNP00996 (Tianjin Museum of Natural History), IVPP V13351 and IVPP V13476 as *Microraptor* sp., and IVPP V13477 as Dromaeosauridae gen. et sp. indet. These specimens provide new information on the morphology and distribution of feathers on non-avian dromaeosaurids. Here we describe the new dromaeosaurid species and in particular the morphology and distribution of feathers on the newly collected dromaeosaurid specimens.

Theropoda Marsh, 1881
Maniraptora Gauthier, 1986
Dromaeosauridae Matthew & Brown, 1922
Microraptor Xu, Zhou & Wang, 2000
Microraptor gui sp. nov.

Etymology. The specific name is in honour of Gu Zhiwei, a distinguished palaeontologist who contributed greatly to the study of Jehol biota.

Material. IVPP V13352 (holotype) and V13320 (referred specimen), both represented by an almost complete skeleton.

Locality and horizon. Dapingfang, Chaoyang County, western Liaoning (30 km southwest of Chaoyang City); Jiufotang Formation¹⁸ (Early Cretaceous).

Diagnosis. Distinguishable from *Microraptor zhaoianus* in having prominent biceps tubercosity on radius, much shorter manual digit I, strongly curved pubis, and bowed tibia.

Description. *Microraptor gui* is a small animal, the holotype being approximately 77 cm in total length (Fig. 1a). Little can be said about the cranial morphology but a tri-radiate postorbital is identifiable. As in *M. zhaoianus*⁸, the basal troodontid *Sinovenator*¹⁵, and the basal oviraptorosaur *Caudipteryx*¹⁹, *M. gui* has a relatively short trunk length, which is 44–50% of hindlimb length (according to the method of ref. 19). The tail is long (Fig. 1a) but has relatively few vertebrae (approximately 26). The middle and posterior caudals are significantly elongate as in other basal dromaeosaurs, basal troodontids and *Archaeopteryx*²⁰. The sternum is a single, flat and large bone (Fig. 1a), different from the condition in other dromaeosaurids^{7,21} where two unfused sternal plates are present. At least seven pairs of slender uncinat processes are present. The anterior uncinat processes cover three ribs and the posteriormost one is short and does not reach the succeeding rib. The fused scapula and coracoid are similar to those of *Sinornithosaurus* and *M. zhaoianus* in the following features: scapula shorter than humerus; glenoid fossa laterally faced; angle between scapula and coracoid less than 90°; and large supracoracoid fenestra present on coracoid²⁰. The forelimb is approximately 2.7 times the femoral length. The ulna is bowed and the radius is much thinner than the ulna (Fig. 1). Metacarpal I is about one-quarter of metacarpal II in length and metacarpal III is slender and bowed laterally. Manual digit II, particularly phalanx II-1, is thick. The pelvis displays the following derived features as in *M. zhaoianus*, *Sinornithosaurus*, *Sinovenator* and basal birds^{5,7,15}: postacetabular process of the ilium tapered; pubis retroverted; and ischium short, with a distally located obturator process and two dorsal processes. The tibia is variably bowed, more so in the referred specimen than the holotype. The pes is similar to that of *M. zhaoianus* in showing a sub-arcotometatarsalian

Table 1 Measurements of femoral length and the length and asymmetry ratio of feathers from the forelimb, hindlimb and tail

Specimen	Femoral length	A distal primary feather	A middle secondary feather	A proximal secondary feather	A distal metatarsus feather	A proximal metatarsus feather	A distal tail feather
IVPP V13352	97	222/1.94* (7)†	95*/1.52 (10)†	81/1.0 (15)†	194/? (2)†	104/1.33 (13)†	120/? (2)†
TNP00996	63	??	??	??	121/1.56* (1)†	72/1.1 (13)†	100/2.1* (1)†
IVPP V13477	72	??	??	??	113/3.08 (2)†	??	??
IVPP V13351	81	??	??	??	190/? (4)†	??	185/? (1)†
IVPP V13320	61	186/1.5* (10)†	??	??	?/2.4 (1)†	130/? (10)†	??
IVPP V13476	94*	??	??	??	175/? (1)†	??	??

Length measurements are given in millimetres. Asymmetry ratio is measured based on the method of ref. 46. Numbers in parentheses indicate the anatomical positions of the feathers (the metatarsal feathers are numbered from distal).

* Incomplete measurements.

† Estimation.

condition and in having slender, strongly curved claws^{8,20}.

Microraptor gui can be unequivocally referred to Dromaeosauridae based on the following derived characters²⁰: extremely elongate prezygapophyses and chevrons; manual phalanges III-1 significantly longer than III-2; specialized pedal digit II; and long metatarsal V. Furthermore it can be referred to *Microraptor* on the basis of the following features²⁰: metacarpal III subequal to metacarpal II in length; extremely short manual phalanx III-2 that is less than one-quarter of manual III-1 length; manual III-3 extremely slender and shorter than III-1 in length, and small distal articulation

of manual III-3 skewed ventrally. However, a few features distinguish it from *M. zhaoianus*. A prominent biceps tubercity is present close to the proximal end of the radius and this feature has not been reported in most other non-avian theropods except a recently described therizinosauroid²². As in most birds, *M. gui* has a proportionately very short manual digit I (metacarpal I + phalanx I-1/metacarpal II length ratio is 0.80–0.84). For comparison, this ratio is 0.97 in *M. zhaoianus* and more than 1.0 in most other non-avian theropods and the basal birds *Archaeopteryx* and *Confuciusornis*. The pubis of *M. gui* is strongly curved (120°), whereas the

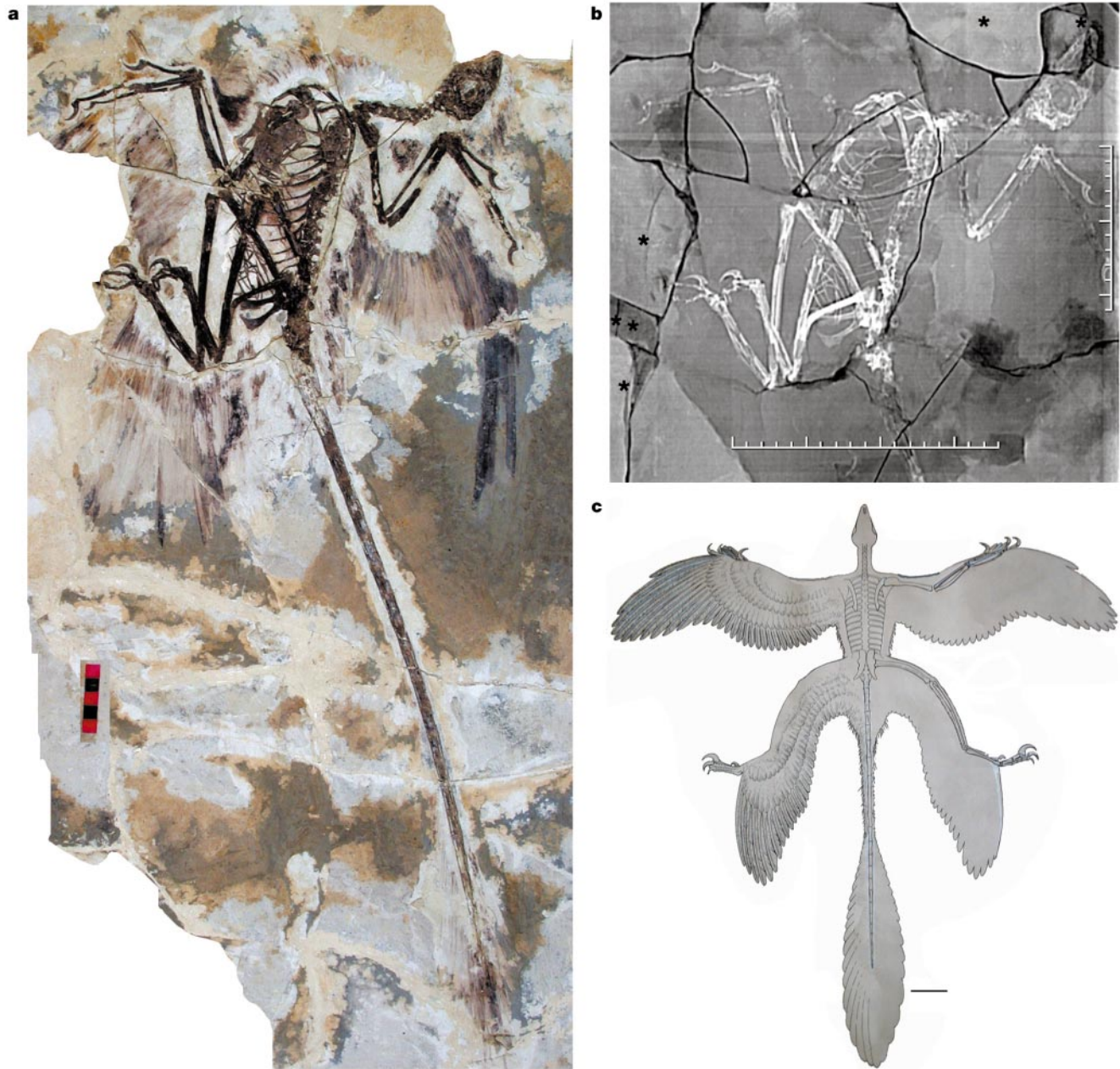


Figure 1 *Microraptor gui*. **a**, Skeleton of *Microraptor gui* (IVPP V13352). Scale bar, 5 cm. **b**, A computerized tomography (CT) image of the major part of the IVPP V13352. Scale bar, 13 cm. Scanning was performed using a CT machine (LightSpeed Qx/i) at an energy level of 140 kV and 250 mA. The images were collected at a size of 800 × 600 pixels. On the basis of comparison of adjacent fracture-face geometries, density of adjacent pieces, and continuity across fractures of bones (see ref. 45), we find a few pieces are unverified or assembled in the wrong position (marked by asterisks). For

example, one small piece containing the anterior end of the skull and a medium-sized piece near the right forelimb preserving some arm feathers are dubious. The latter is actually from the counter slab. However, the CT information suggests that most pieces lie together in their natural relationships, including pieces containing the forelimb, hindlimb and associated feather impressions. This is concordant with microscopic observations. **c**, A reconstruction of *M. gui* showing the morphology and distribution of the pennaceous feathers. Scale bar, 6 cm.

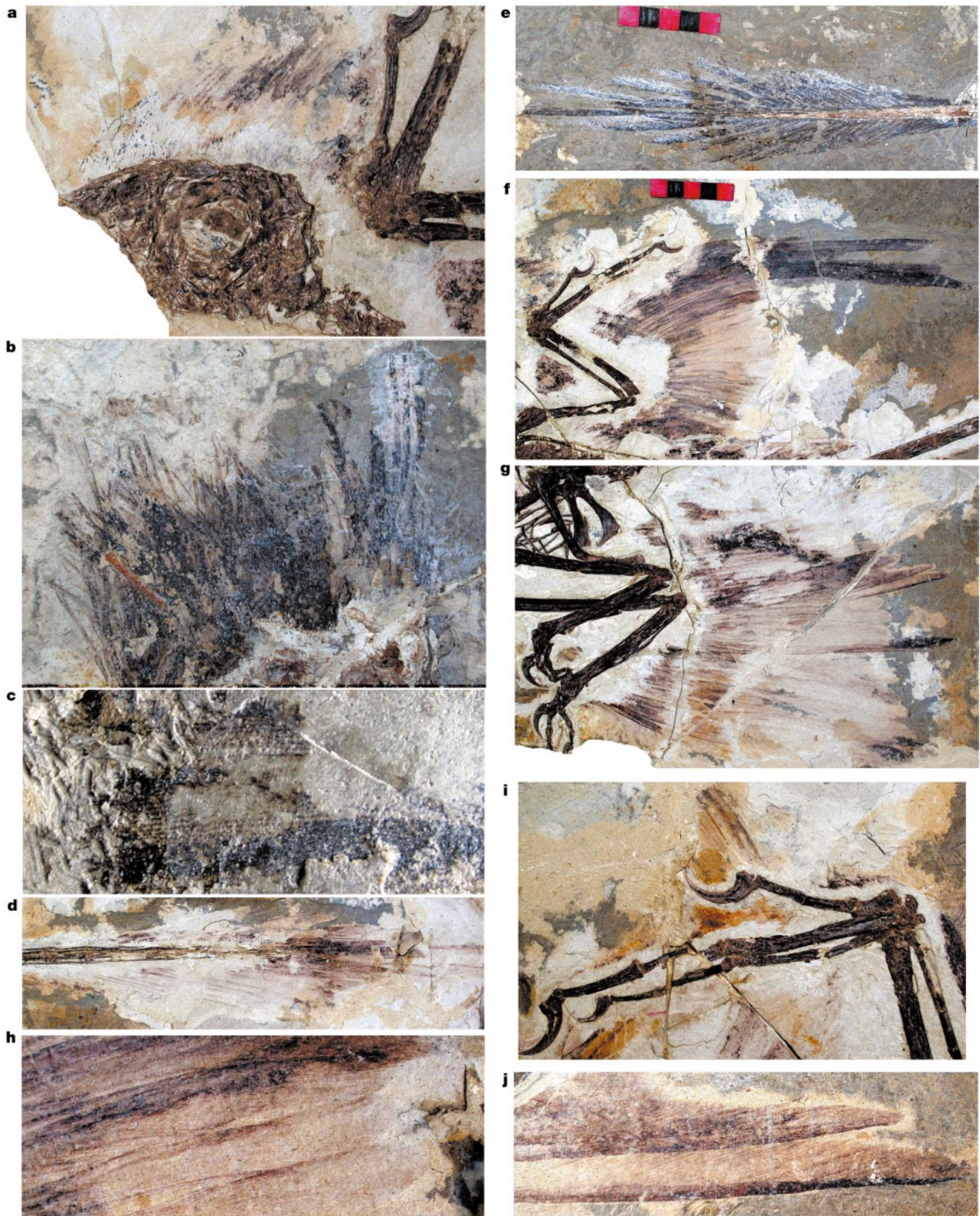


Figure 2 Feathers of IVPP V13352 and TNP00996. Feathers attached to the skull (a), the tail (d), the forelimb (f), the manual digit I (i), and the hindlimb (g) of IVPP V13352, and to the skull (b) and the tail (e) of TNP00996; close-up of the skull feathers of TNP 00996 (c), and of secondaries (h) and large pennaceous feathers on distal metatarsus (j) of IVPP

V13352. Note the pennaceous feathers attached to the digit (i) that might be a precursor to the alula. This is concordant with the fact that *M. gui* has a short manual digit I, because the alula is often associated with a reduced alular digit except in *Protopteryx*²⁴. Scale bar, 5 cm.

pubis is relatively straight in most other non-avian theropods, including *M. zhaoianus*. The other distinctive feature of *M. gui* is the bowed tibia whereas in most other theropods, it is straight. These features suggest that *M. gui* is a new species.

Integument. Information on the integument is based on the holotype (IVPP V13352) and referred specimen (V13320) of *M. gui* as well as TNP00996, IVPP V13351, IVPP V13476 and IVPP V13477 (Table 1). IVPP V13476 and IVPP V13477 were collected from Shangheshou (3 km northwest of Chaoyang City, Liaoning) and the other specimens are from Dapingfang. The integumentary remains are best preserved in IVPP V13352 and TNP00996, in which they are well preserved around the whole skeleton. The integument displays two types: plumulaceous and pennaceous feathers. The body is covered by plumulaceous feathers that are about 25–30 mm long. The feathers attached to the skull roof are up to 40 mm long in IVPP V13352 (Fig. 2a), and in TNP00996 they are proportionately even longer (Fig. 2b). Some feathers on the head display well-organized pennaceous vanes (Fig. 2c). These feathers are most probably functionally related to display, as in some modern birds such as *Pitheophaga jefferyi*. Large pennaceous feathers are attached to the distal tail, forelimb and hindlimb (Figs 1a, c and 2d–g). The remiges (wing feathers) are preserved in a pattern similar to those of modern birds (Fig. 2f). The

primaries (approximately 12 in number) are significantly longer than the secondaries (approximately 18 in number); the most distal primaries are more or less parallel to the manus, and the others are at angle to the manus, with the angle increasing from distal to proximal. The longest primaries (incompletely preserved) are 2.7 times as long as the humerus or 2.3 times as long as the femur. Some primaries on the holotype display asymmetry, with the leading vane much narrower than the trailing vane. The secondaries are longer than the humerus and more or less perpendicular to the ulna. The proximal ones have symmetrical vanes and the distal ones display weak asymmetry (Fig. 2h). The presence of a few relatively small feathers attached to the manual digit I (Fig. 2i) on the holotype is noteworthy. These display well-organized pennaceous vanes, and might be the precursor to the alula, which is associated with flight control and which is present in most birds other than *Archaeopteryx* and *Confuciusornis*^{23,24}. Coverts are present and appear to be variable in size. Some coverts can be identified as under-wing coverts. The presence of remex-like feathers along the hindlimbs is most unusual. The leg feathers are arranged in a pattern similar to the arm feathers (Fig. 2g). There are at least 14 large pennaceous feathers attached to the metatarsus; different from the primaries, they are more or less perpendicular to the metatarsus. The distal feathers have asymmetrical vanes (Fig. 2j) and the proximal feathers



Figure 3 Feathers showing the asymmetrical vanes. Pennaceous feathers attached to the left metatarsus of IVPP V13477 (**a**) and right metatarsus of IVPP V13320 (**b**). **c**, Close-up of a distal pennaceous feather on the left metatarsus of IVPP V13477. **d**, Close-up of a

distal pennaceous feather on the left metatarsus of TNP00996. **e**, Close-up of some tail feathers of TNP00996. **f**, Close-up of some distal primaries of IVPP V13352.

are shorter in length and have symmetrical vanes. Coverts are also seen attached to the metatarsus and seem to be denser than those on the forelimb. Pennaceous feathers are also present on the tibia, and they appear to be shorter than the feathers on the metatarsus. The longest leg feathers (incompletely preserved) are more than twice femoral length and located close to the distal end of the metatarsus. The rectrices (tail feathers) begin with the 15th to 18th caudal through the end of the tail. The rectrices in basal dromaeosaurs extend less proximally (Figs 1c and 2d, e) on the tail than in *Archaeopteryx* but more proximally than in *Caudipteryx*⁶. The most distal rectrices are more than 120 mm in length on the holotype, and in TNP00996 the longest rectrices are about 105 mm relative to a 290-mm-long tail.

Discussion. Recent discoveries suggest that pennaceous feathers are present on Dromaeosauridae¹⁷, a non-avian theropod group, and our observations on the newly collected specimens provide new information that is important for understanding the transition towards birds.

The most unusual feature is the attachment of pennaceous feathers to the whole length of the metatarsus (Figs 1a, c, 2g and 3a). They are long and some have asymmetrical vanes like flight feathers (Fig. 3a–d). We exclude the possibility that these are preservation artefacts because we observed this feature in all six specimens in the present study, most of which are represented by well-articulated skeletons. Pennaceous feathers are also associated with the tibia and femur¹⁹ and they display symmetrical vanes. In general, the leg feathers are arranged in a pattern similar to wing feathers in modern birds, suggesting the presence of a hindlimb wing. Although there is no modern analogue, our observations are concordant with some early hypotheses^{25,26} that there is a tetrapteryx stage in bird evolution.

Recent work shows that basal dromaeosaurs closely resemble *Archaeopteryx* in flight apparatus^{7,8,20,27} (except for some differences in limb proportion, such as the longer arm in the latter taxon), and thus dromaeosaurs were pre-adapted for flight. The new discoveries provide further information on soft tissues and greatly improve our knowledge of these close relatives of birds. Asymmetrical pennaceous feathers are suggested to have the aerodynamic function necessary for flight²⁸. The asymmetry is present not only on the forelimb and tail feathers, but also the hindlimb feathers of basal dromaeosaurs (Fig. 3c–f). The forelimb and the leg feathers would make a perfect aerofoil together, analogous to the patagium in bats or gliding animals. These features together suggest that basal dromaeosaurids probably could glide, representing an intermediate stage between the flightless non-avian theropods and the volant avialans. Apparently some non-avian theropods evolved large and highly specialized pennaceous feathers on the leg for aerodynamic function; these features were later reduced and lost in birds, which depend completely on forewings for flight.

The metatarsus feathers are inconsistent with the suggestions that basal dromaeosaurs are cursorial animals^{1,29} because such long feathers on the feet would be a hindrance for a small cursorial animal. It is unlikely that a small dromaeosaur could run fast with such an unusual integument and this provides negative evidence for the ground-up hypothesis for the origin of avian flight^{29–31}. Some recent osteological studies suggested that non-avian theropods^{8,20,32} and basal birds^{33–36} acquired arboreal capabilities, which were later improved in more derived birds^{20,37,38}. Combined with the new information from the integument, we suggest that basal dromaeosaurs were arboreal animals, and that the ancestor of birds first learned to glide by taking advantage of gravity before flapping flight was acquired in birds^{32,39–44}.

Notes on the specimens. Of the six specimens in the present study, IVPP V13476 was collected by the Liaoxi expedition team of the IVPP in 2001, IVPP V13352, V13320, V13477 and V13351 were purchased by the IVPP during the field seasons of 2001 and 2002, and TNP00996 was purchased by Tianjin Museum of Natural

History in 2002. Some counter slabs of IVPP V13352 and V13477 are available. IVPP V13476 has been prepared at the IVPP. Although much of the preparation was done before we obtained the other specimens, some preparation of IVPP V13352 and V13477 was done at the IVPP. The preparation at the IVPP exposed some parts of the integument, including hindlimb feathers that were covered by the matrix. We observed that there are some pieces of blocks mistakenly glued to the specimens; however, we excluded all the dubious parts from the study (Fig. 1b). We carefully examined the specimens under the microscope and with high-resolution X-ray computerized tomography (CT) to test the authenticity of one of the studied specimens⁴⁵ (IVPP V13352) and can guarantee the accuracy of the information that we provide in this study. □

Received 5 August; accepted 29 November 2002; doi:10.1038/nature01342.

- Ostrom, J. H. Osteology of *Deinonychus antirrhopus*, an unusual theropod dinosaur from the Lower Cretaceous of Montana. *Peabody Mus. Nat. Hist. Bull.* **30**, 1–165 (1969).
- Currie, P. J. Cranial anatomy of *Stenonychosaurus inequalis* (Saurischia, Theropoda) and its bearing on the origin of birds. *Can. J. Earth Sci.* **22**, 1643–1658 (1985).
- Norell, M. A., Clark, J. M., Chiappe, L. M. & Dashzeveg, D. A nesting dinosaur. *Nature* **378**, 774–776 (1995).
- Novas, F. E. & Puerta, P. New evidence concerning avian origins from the Late Cretaceous of Patagonia. *Nature* **387**, 390–392 (1998).
- Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279**, 1915–1919 (1998).
- Ji, Q., Currie, P. J., Norell, M. A. & Ji, S.-A. Two feathered dinosaurs from northeastern China. *Nature* **393**, 753–761 (1998).
- Xu, X., Wang, X.-L. & Wu, X.-C. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401**, 262–266 (1999).
- Xu, X., Zhou, Z.-H. & Wang, X.-L. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2000).
- Zhou, Z.-H. & Zhang, F.-C. A long-tailed, seed-eating bird from the Early Cretaceous of China. *Nature* **418**, 405–409 (2002).
- Gauthier, J. A. Saurischian monophyly and the origin of birds. *Mem. Calif. Acad. Sci.* **8**, 1–55 (1986).
- Holtz, T. R. Jr The phylogenetic position of the Tyrannosauridae: implications for theropod systematics. *J. Paleontol.* **68**, 1100–1117 (1994).
- Holtz, T. R. Jr A new phylogeny of the carnivorous dinosaurs. *Gaia* **15**, 5–61 (2000).
- Sereno, P. C. The evolution of dinosaurs. *Science* **284**, 2137–2147 (1999).
- Norell, M. A., Clark, J. M. & Makovicky, P. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 49–67 (Peabody Mus. Nat. Hist., Yale Univ., New Haven, 2001).
- Xu, X., Norell, M. A., Wang, X.-L., Makovicky, P. & Wu, X.-C. A basal troodontid from the Early Cretaceous of China. *Nature* **415**, 780–784 (2002).
- Makovicky, P. & Sues, H.-D. Anatomy and phylogenetic relationships of the theropod dinosaur *Microvenator celer* from the Lower Cretaceous of Montana. *Am. Mus. Novit.* **3240**, 1–27 (1998).
- Norell, M. A. *et al.* 'Modern' feathers on a non-avian dinosaur. *Nature* **416**, 36–37 (2002).
- Smith, P. E. *et al.* Dates and rates in ancient lakes: ⁴⁰Ar–³⁹Ar evidence for an early Cretaceous for the Jehol Group, northeast China. *Can. J. Earth Sci.* **32**, 1426–1431 (1995).
- Jones, T. D., Farlow, J. O., Ruben, J. A., Henderson, D. M. & Hillenius, W. J. Cursoriality in bipedal archosaurs. *Nature* **406**, 716–718 (2000).
- Xu, X. *Deinonychosaurian Fossils from the Jehol Group of Western Liaoning and the Coelurosaurian Evolution*. PhD dissertation, Chinese Academy of Sciences (2002).
- Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaur skeleton: information from a new specimen. *Am. Mus. Novit.* **3215**, 1–28 (1997).
- Zhang, X.-H. *et al.* A long-necked therizinosaurid dinosaur from the Upper Cretaceous Iren Dabasu Formation of Nei Mongol, People's Republic of China. *Vertebrata Palasiatica* **39**, 282–290 (2001).
- Sanz, J. L. *et al.* An Early Cretaceous bird from Spain and its implications for the evolution of avian flight. *Nature* **382**, 442–445 (1996).
- Zhang, F.-C. & Zhou, Z.-H. A primitive enantiornithine bird and the origin of feathers. *Science* **290**, 1955–1959 (2000).
- Beebe, C. W. *A Tetrapteryx Stage in the Ancestry of Birds* (Zoologica, New York, 1915).
- Heilmann, G. *The Origin of Birds* (Witherby, London, 1927).
- Burnham, D. A. *et al.* Remarkable new birdlike dinosaur (Theropoda: Maniraptora) from the Upper Cretaceous of Montana. *Univ. Kansas Paleont. Contr. (new series)* **13**, 1–14 (2000).
- Feduccia, A. & Tordoff, H. B. Feathers of *Archaeopteryx*: asymmetric vanes indicate aerodynamic function. *Science* **203**, 1021–1022 (1979).
- Ostrom, J. H. The cursorial origin of avian flight. *Mem. Calif. Acad. Sci.* **8**, 73–81 (1986).
- Padian, K. & Chiappe, L. M. The origin and early evolution of birds. *Biol. Rev.* **73**, 1–42 (1998).
- Burgers, P. & Chiappe, L. M. The wing of *Archaeopteryx* as a primary thrust generator. *Nature* **399**, 60–62 (1999).
- Chatterjee, S. *The Rise of Birds* (Johns Hopkins Univ. Press, Baltimore, 1997).
- Yalden, D. W. Climbing *Archaeopteryx*. *Archaeopteryx* **15**, 107–108 (1997).
- Chiappe, L. M., Ji, S.-A., Ji, Q. & Norell, M. A. Anatomy and systematics of the Confuciusornithidae (Theropoda: Aves) from the late Mesozoic of Northeastern China. *Bull. Am. Mus. Nat. Hist.* **242**, 1–89 (1999).
- Hopson, J. A. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 211–236 (Peabody Mus. Nat. Hist., Yale Univ., New Haven, 2001).
- Zhou, Z.-H. & Farlow, J. O. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 237–254 (Peabody Mus. Nat. Hist., Yale Univ., New Haven, 2001).
- Sereno, P. & Rao, C. Early evolution of avian flight and perching: new evidence from the Early Cretaceous of China. *Science* **255**, 845–848 (1992).
- Arnold, E. N. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall,

- L. F.) 195–210 (Peabody Mus. Nat. Hist., Yale Univ., New Haven, 2001).
39. Norberg, U. M. Evolution of vertebrate flight: an aerodynamic model for the transition from gliding to active flight. *Am. Nat.* **126**, 303–327 (1985).
 40. Norberg, U. M. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 293–302 (Freunde des Jura-Museums Eichstatt, Eichstatt, 1985).
 41. Norberg, U. M. *Vertebrate Flight* (Springer, Berlin, 1990).
 42. Rayner, J. M. V. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 279–288 (Freunde des Jura-Museums Eichstatt, Eichstatt, 1985).
 43. Rayner, J. M. V. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 289–292 (Freunde des Jura-Museums Eichstatt, Eichstatt, 1985).
 44. Bock, W. J. The arboreal origin of avian flight. *Mem. Calif. Acad. Sci.* **8**, 57–72 (1986).
 45. Rowe, T. *et al.* The Archaeoraptor forgery. *Nature* **410**, 539–540 (2001).
 46. Speakman, J. R. & Thomson, S. C. Flight capabilities of *Archaeopteryx*. *Nature* **370**, 524 (1994).

Acknowledgements We thank L. Witmer and A. Milner for their suggestions and comments, M. Norell, X.-J. Ni, J. Liu, J. Clarke and P. Sereno for discussions, Y.-L. Huo, Y.-T. Li and H.-J. Wang for preparing the specimens, R.-S. Li for drawings, and Z.-G. Sun and B. An for help with CT scanning. Thanks also go to members of the Liaoxi expedition team of the IVPP for fieldwork. This work was supported by grants from the National Natural Science Foundation of China, the National Geographic Society, Special Funds for Major State Basic Research Projects of China, and the Chinese Academy of Sciences.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.X. (e-mail: xing_xu@sina.com).

Spectral signature of cosmological infall of gas around the first quasars

Rennan Barkana* & Abraham Loeb†‡

* School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel

† Astronomy Department, Harvard University, Cambridge, Massachusetts 02138, USA

Recent observations have shown that, only a billion years after the Big Bang, the Universe was already lit up by bright quasars¹ fuelled by the infall of gas onto supermassive black holes. The masses of these early black holes are inferred from their luminosities to be $>10^9$ solar masses ($M_\odot$), which is a difficult theoretical challenge to explain. Like nearby quasars, the early objects could have formed in the central cores of massive host galaxies. The formation of these hosts could be explained if, like local large galaxies, they were assembled gravitationally inside massive ($>10^{12} M_\odot$) haloes of dark matter². There has hitherto been no observational evidence for the presence of these massive hosts or their surrounding haloes. Here we show that the cosmic gas surrounding each halo must respond to its strong gravitational pull, where absorption by the infalling hydrogen produces a distinct spectral signature. That signature can be seen in recent data^{3,4}.

We model the effect of resonant Lyman α absorption by infalling gas on the quasar light (Fig. 1). For each quasar we consider the history of the formation of its host halo from an initial positive overdensity of dark matter. For the initial surrounding density profile at high redshift we adopt the typical profile expected around the dense region that collapses to form the halo⁵. This profile, as well as the approximation of a spherical geometry, are particularly satisfactory for the very rare and massive haloes under consideration. We calculate gas infall down to the radius of the accretion shock, and neglect any Ly α absorption due to the post-shock gas. The hot ($\approx 10^7$ K) post-shock gas should be fully ionized by collisions. Part of it is expected to subsequently cool and collapse onto the galactic disk, but Compton heating by the quasar should keep the virialized gas hotter than about 10^6 K. Even if a thin cold shell of shocked gas remains, it will not change the basic pattern produced by infalling gas because the post-shock gas no longer has a high infall velocity.

In order to predict the Ly α absorption around a quasar we must estimate the mass of its host halo. A tight correlation has been measured in local galaxies between the mass of the central black hole and the bulge velocity dispersion^{6,7}. This relation also fits all existing data on the luminosity function of high-redshift quasars within a simple model⁸ in which quasar emission is assumed to be triggered by mergers during hierarchical galaxy formation. We use the best-fit⁶⁻⁸ relation, in which the black hole mass (M_{BH}) in multiples of $10^8 M_\odot$ is related to the circular velocity (V) at the halo boundary in multiples of 300 km s^{-1} by $M_{\text{BH}} = 1.5V^5$. For the typical quasar continuum spectrum, we adopt a power-law spectral flux shape of $F_\nu \propto \nu^{-0.44}$ in the rest-frame range $1,190\text{--}5,000 \text{ \AA}$, based on the Sloan Digital Sky Survey (SDSS) composite spectrum⁹, and $F_\nu \propto \nu^{-1.57}$ at $500\text{--}1,190 \text{ \AA}$, using the composite quasar spectrum from the Hubble Space Telescope¹⁰. On the basis of observations in soft X-rays¹¹, we extend this power-law towards short wavelengths.

We assume that the brightest quasars shine at their Eddington luminosity, and we note that for the SDSS composite spectrum⁹, the total luminosity above $1,190 \text{ \AA}$ equals 1.6 times the total continuum luminosity at $1,190\text{--}5,000 \text{ \AA}$. Thus, ionizing photons stream out of the quasar's galactic host at the rate $\dot{N} = 1.04 \times 10^{56} M_{\text{BH}} \text{ s}^{-1}$. We

infer the black hole mass using the observed continuum at $1,350 \text{ \AA}$, with the conversion $F_\nu = 1.74 \times 10^{30} M_{\text{BH}} \text{ erg s}^{-1} \text{ Hz}^{-1}$. Given that helium is doubly ionized by the quasar, the frequency-averaged photoionization cross-section of hydrogen is, for our template spectrum, $\sigma_{\text{H}} = 2.3 \times 10^{-18} \text{ cm}^2$. The double ionization of helium increases the recombination rate owing to the extra electrons, and also produces a characteristic gas temperature of around $1.5 \times 10^4 \text{ K}$ in regions that have already been reionized by a softer ionizing background¹².

Only a limited number of published spectra are currently available for a clear test of our predictions. First, only the brightest quasars, which reside in the most massive haloes, produce strong infall over a large surrounding region. The infalling gas density scales as $(1+z)^3$ and the shock radius scales as $(1+z)^{-1}$, so that the absorption optical depth (see below) scales roughly as $(1+z)^4$ and the absorption feature is much weaker at low redshift. Even at high redshifts, measuring the detailed Ly α line profile is possible only in a high-resolution spectrum with an extremely high signal-to-noise ratio.

Figure 2 shows two particularly high-quality spectra and compares them with our model predictions. Both spectra show our predicted double-peak pattern and disagree with the single-peaked profile predicted by previous models that ignored infall. In particular,

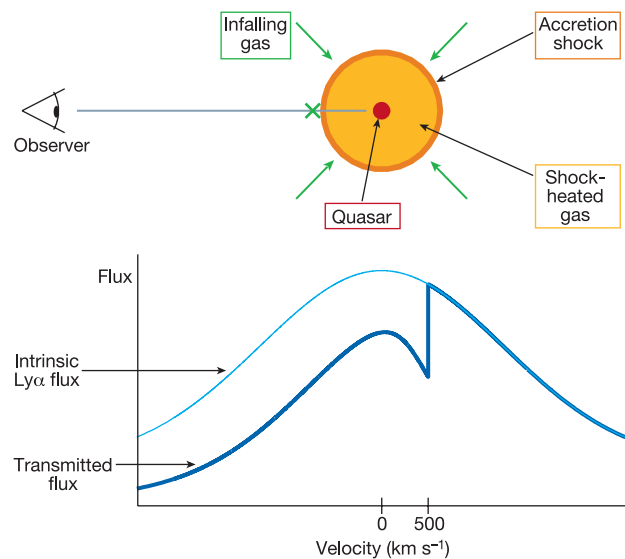


Figure 1 Schematic illustration of how infall produces a unique spectral signature. A quasar forms inside a galaxy that lies at the centre of a massive dark matter halo. A large volume of gas responds to the strong gravitational pull and falls towards the massive halo. As infalling gas impacts on the galactic gas, a strong accretion shock forms. The intrinsic quasar Lyman α emission is partially absorbed by the infalling pre-shock gas. In particular, a sharp flux drop is caused by gas (marked with a cross) that is about to hit the accretion shock. This gas is falling towards the quasar at 500 km s^{-1} in this sketch. We calculate spherically symmetric infall¹⁷ and set the accretion shock radius to 1.15 times the halo boundary¹⁸, although our results are not altered substantially as long as the shock radius is close to the halo virial radius. Three-dimensional hydrodynamic simulations show that the most massive haloes at any time in the universe are indeed surrounded by strong, quasi-spherical accretion shocks^{19,20}. Preliminary evidence has been found for such shocks around nearby galaxy clusters^{21,22}. We model resonant Ly α absorption by intergalactic hydrogen²³ that is partially ionized by the radiation produced by the quasar²⁴. In addition to the overall infall pattern⁵ we include a realistic distribution of gas density fluctuations²⁵. For the distribution of gas clumps we adopt an analytical fit to numerical simulations²⁶. We assume that the clumps are optically thin, and find the neutral fraction separately for each clumping density based on ionization equilibrium with the quasar ionizing flux. We then calculate the mean Ly α transmission averaged over the clump distribution. Note that the absorbed Ly α photons are re-emitted from a large Ly α halo^{27,28} that is too faint to significantly affect current observations.

‡ Present address: Institute for Advanced Study, Princeton, New Jersey 08540, USA.

models that assume a pure Hubble flow firmly predict no absorption at all at positive velocities and only a gradual decline in the transmitted flux toward negative velocities. This slow decline is due to the fact that in these models the gas closest to the quasar produces no absorption because it is fully ionized by the strong ionizing intensity of the quasar. More distant gas recedes along with the universal expansion and absorbs only at negative velocities. Our model, in contrast, firmly predicts a sharp flux cutoff located at a positive velocity. This flux drop corresponds to absorption by gas just outside the accretion shock, and the positive velocity of the cutoff corresponds to the infall velocity. The strong absorption caused by this gas, despite exposure to the ionizing flux from the quasar, is due to the high density of the infalling gas. This gas is expected to be about 10–30 times denser than the cosmic mean after having fallen toward the quasar host halo over the history of its formation.

The infall velocity is proportional to the halo circular velocity and thus² to $M^{1/3}(1+z)^{3/2}$, in terms of the halo mass M and the redshift z . The actual physical distance of the accretion shock in the model is 86 kpc and 80 kpc for the quasars at $z = 4.795$ and $z = 6.28$, respectively. Our model with infall also fits approximately the secondary peak that is observed on the blue side of the main peak (that is, at lower velocity). This blue peak corresponds closely to the point of weakest absorption by the infalling gas, although the profile of the intrinsic quasar emission also affects the precise location of this peak. In this region we do not expect to fit the flux profile in full detail; our model averages over all lines of sight and possible quasar positions, but observable random fluctuations around our predicted mean are expected in each specific spectrum.

However, once we average over the density fluctuations, the prediction of a second peak rather than a smooth fall-off is generic and insensitive to the detailed model assumptions. At a distance R from the quasar the resonant optical depth depends on $\rho^2 R^2$, where ρ is the density including infall; one factor of ρ comes from the total gas density, the second comes from the H I fraction which increases with ρ owing to recombinations, and the R -dependence results from the R^{-2} decline of the ionizing intensity of the quasar. Thus, a second peak should appear as long as infall produces a density profile falling off faster than $1/R$ (our model predicts $R^{-3/2}$) before asymptotically tending to the cosmic mean value of unity. We note that in our model the position of the blue peak corresponds to absorption by gas at $R = 0.45$ Mpc ($z = 4.795$) or 0.41 Mpc ($z = 6.28$).

Even in each particular spectrum, the more distant region where the flux drops toward zero can be modelled much more robustly. In particular, significant flux is observed at $\Delta V = -2,000$ km s⁻¹ ($z = 4.795$) and $\Delta V = -3,000$ km s⁻¹ ($z = 6.28$), respectively. These relatively distant regions are only weakly affected by infall and the observed positions translate to a distance from the quasar of 3.8 Mpc ($z = 4.795$) or 4.2 Mpc ($z = 6.28$). In models that do not include a clump distribution, the optical depth at these positions is ≥ 3 , which means that the observed flux requires an intrinsic unabsorbed flux that is 20 times greater. This is clearly impossible, regardless of any uncertainties about the intrinsic line shape and the quasar continuum level. To explain the observed flux, the ionizing intensity of the quasar as determined by our template spectrum from the observed continuum would have to be too low by a factor greater than 2. However, our full model accounts for the fact that part of the cosmic gas falls into dense sheets and filaments and leaves the rest with a density below the cosmic mean. The resonant Ly α absorption is made up of the separate contributions of gas elements at a variety of overdensities, and because gas in low-density regions absorbs very weakly, clumping actually increases the mean transmission. Thus, our model naturally accounts for the flux observed far from the quasar, with no change required in the quasar spectrum.

Our conclusions are insensitive to the question of whether the

H II region of the highest redshift quasar is surrounded by a region of neutral hydrogen (owing to the fact that the universe had not been fully reionized by $z = 6.28$ (ref. 13)) or not¹⁴; a distant neutral region would only add on the intergalactic medium damping wing¹⁵, which produces a smooth, gradual suppression that should not alter the basic double-peak pattern. We note that the quasar may possess a velocity offset relative to Hubble flow owing, for example, to a violent galactic merger that had originally activated the quasar. However, the close fit that we find between the predicted accretion shock position and the observed flux drop is evidence against the presence of a large velocity offset in the two quasars we have considered. We also note that if the observed absorption pattern were due to dense gas clouds within the galaxy (for example, in the region feeding the central black hole), then the absorbing gas would be expected to contain heavy elements such as carbon, oxygen and nitrogen. This gas would then be expected to absorb other emission lines of the quasar in addition to Ly α . Such associated absorption is absent in the $z = 4.795$ quasar which has several emission lines with

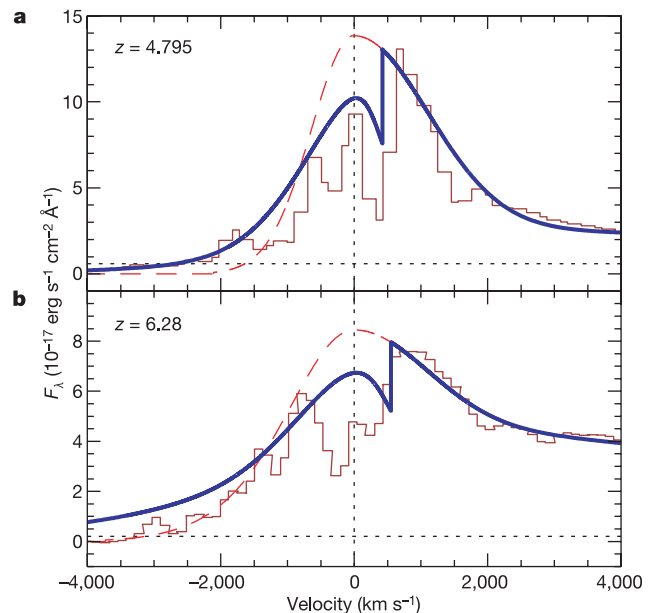


Figure 2 Comparison between models of cosmological infall and observed quasar spectra. **a**, The redshift 4.795 ± 0.004 quasar SDSS1122–0229 (ref. 3); on the basis of observed redshift and continuum level our model implies a $4.6 \times 10^8 M_{\odot}$ black hole residing in a $2.5 \times 10^{12} M_{\odot}$ host halo. **b**, The redshift 6.28 ± 0.02 quasar SDSS1030 + 0524 (ref. 4), for which our model implies a $1.9 \times 10^9 M_{\odot}$ black hole residing in a $4.0 \times 10^{12} M_{\odot}$ host halo (we do not use a second spectral observation of this source²⁹ because it appears to have a significantly lower signal-to-noise ratio). In **a** and **b**, the histogram shows the observed spectrum, the dashed line shows previous models that assume a uniform expanding universe, and the solid line shows our model which includes cosmological infall as well as a realistic distribution of gas clumps. Note that the velocity is measured relative to the quasar, where negative velocity means motion towards us. In **a** and **b**, the vertical dotted line shows the position of the Ly α wavelength at the source redshift, and the horizontal dotted line shows the flux level of the highest transmission peaks seen in parts of the spectrum corresponding to the average intergalactic medium (that is, at velocities more negative than $-4,000$ km s⁻¹). We assume an intrinsic emission line given by a sum of two gaussian components, a form which best fits the line shape of most quasars at low redshift³⁰; we fix the parameters for each quasar on the basis of the unabsorbed part of the Ly α line at velocities above 500 km s⁻¹. With this approach the models do not include any free parameters. Particular quasars are expected to show fluctuations around our predicted absorption profile, because our model averages over random lines of sight and density fluctuations. Throughout this paper we assume the standard cosmological parameters $\Omega_m = 0.3$, $\Omega_{\Lambda} = 0.7$, $\Omega_b = 0.05$, $H_0 = 70$ km s⁻¹ Mpc⁻¹ and $n = 1$. F_{λ} is the total energy from the quasar crossing a unit area per second per unit wavelength.

well-measured profiles³, suggesting that the absorbing gas has a near-primordial composition as expected for intergalactic gas.

Our models provide direct evidence that two characteristic properties of quasars at low redshift are also applicable to bright quasars in the early Universe. These properties include the quasar spectral template, which determines the ionizing intensity of the quasar, and the relation between black hole mass and halo velocity dispersion, which we have used to determine the host halo mass. Both observed spectra show a blue peak of about 75% of the height of the red (positive velocity) peak, and this is roughly matched by the models. However, if we were to increase the ionizing intensity by an order of magnitude, then we would predict a blue peak at least of equal height to the other peak. If, instead, we decreased the ionizing intensity by an order of magnitude, then the resulting blue peak would be under 50% of the height of the red peak and the transmitted flux would decrease to zero toward negative velocities much faster than is observed. Similarly, if we varied the assumed halo mass by more than an order of magnitude then the resulting absorption profile in each quasar would disagree with the data. High-redshift quasars could in principle be much fainter intrinsically than they appear, if they are magnified by gravitational lensing¹⁶; our limits on the ionizing intensity, however, suggest that the two quasars we have modelled cannot be magnified by a factor of more than about ten.

We can also estimate from the data the total gas infall rates into these massive galaxies. The positions of the accretion shocks imply, in our models, infall velocities of 400–550 km s⁻¹ and shock radii of 80–90 kpc. Gas at this radius is expected to have a density of about 20 times the cosmic mean density⁵, so we obtain accretion rates of 1,300 M_☉ yr⁻¹ ($z = 4.795$) and 2,900 M_☉ yr⁻¹ ($z = 6.28$), respectively. At these rates, the host galaxies of these two quasars could have been assembled in $(2-3) \times 10^8$ yr, consistent with the 9×10^8 yr age of the Universe at $z = 6.28$. Future comparison of our model to the average Ly α absorption profile of a statistical sample of bright, early quasars with similar luminosities and redshifts should allow us to fit the details of the absorption spectrum and refine our quantitative conclusions. □

Received 25 September; accepted 26 November 2002; doi:10.1038/nature01330.

1. Fan, X. *et al.* Survey of $z > 5.8$ quasars in the Sloan digital sky survey. I. Discovery of three new quasars and the spatial density of luminous quasars at $z \sim 6$. *Astron. J.* **122**, 2833–2849 (2001).
2. Barkana, R. & Loeb, A. In the beginning: the first sources of light and the reionization of the universe. *Phys. Rep.* **349**, 125–238 (2001).
3. Zheng, W. *et al.* Five high-redshift quasars discovered in commissioning imaging data of the Sloan Digital Sky Survey. *Astron. J.* **120**, 1607–1611 (2000).
4. Becker, R. H. *et al.* Evidence for reionization at $z \sim 6$: Detection of a Gunn-Peterson trough in a $z = 6.28$ quasar. *Astron. J.* **122**, 2850–2857 (2001).
5. Loeb, A. & Eisenstein, D. J. Probing early clustering with Ly α absorption lines beyond the quasar redshift. *Astrophys. J.* **448**, 17–26 (1995).
6. Ferrara, A. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539**, L9–L12 (2000).
7. Tremaine, S. *et al.* The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574**, 740–753 (2002).
8. Wyithe, J. S. B., Loeb, A. A physical model for the luminosity function of high-redshift quasars. *Astrophys. J.* (in the press).
9. Vanden Berk, D. E. *et al.* Composite quasar spectra from the Sloan Digital Sky Survey. *Astron. J.* **122**, 549–564 (2001).
10. Telfer, R. C., Zheng, W., Kriss, G. A. & Davidsen, A. F. The rest-frame extreme-ultraviolet spectral properties of quasi-stellar objects. *Astrophys. J.* **565**, 773–785 (2002).
11. Yuan, W., Brinkmann, W., Siebert, J. & Voges, W. Broad band energy distribution of ROSAT detected quasars. II. Radio-quiet objects. *Astron. Astrophys.* **330**, 108–122 (1998).
12. Abel, T. & Haehnelt, M. G. Radiative transfer effects during photoheating of the intergalactic medium. *Astrophys. J.* **520**, L13–L16 (1999).
13. Fan, X. *et al.* Evolution of the ionizing background and the epoch of reionization from the spectra of $z \sim 6$ quasars. *Astron. J.* **123**, 1247–1257 (2002).
14. Barkana, R. Did the universe reionize at redshift six? *New Astron.* **7**, 85–100 (2002).
15. Miralda-Escudé, J. Reionization of the intergalactic medium and the damping wing of the Gunn-Peterson trough. *Astrophys. J.* **501**, 15–22 (1998).
16. Wyithe, J. S. B. & Loeb, A. Magnification of light from many distant quasars by gravitational lenses. *Nature* **417**, 923–925 (2002).
17. Gunn, J. E. & Gott, J. R. On the infall of matter into clusters of galaxies and some effects on their evolution. *Astrophys. J.* **176**, 1–19 (1972).
18. Bertschinger, E. Self-similar secondary infall and accretion in an Einstein-de Sitter universe. *Astrophys. J. Suppl.* **58**, 39–65 (1985).

19. Keshet, U., Waxman, E., Loeb, A., Springel, V., Hernquist, L. Gamma-rays from intergalactic shocks. *Astrophys. J.* (in the press).
20. Abel, T., Bryan, G. L. & Norman, M. L. The formation of the first star in the universe. *Science* **295**, 93–98 (2002).
21. Scharf, C. A., Mukherjee, R. A statistical detection of gamma-ray emission from galaxy clusters: implications for the gamma-ray background and structure formation. *Astrophys. J.* **580**, 154–163 (2002).
22. Loeb, A. & Waxman, E. Cosmic γ -ray background from structure formation in the intergalactic medium. *Nature* **405**, 156–158 (2000).
23. Gunn, J. E. & Peterson, B. A. On the density of neutral hydrogen in intergalactic space. *Astrophys. J.* **142**, 1633–1641 (1965).
24. Bajlik, S., Duncan, R. C. & Ostriker, J. P. Quasar ionization of Ly α clouds—the proximity effect, a probe of the ultraviolet background at high redshift. *Astrophys. J.* **327**, 570–583 (1988).
25. Haiman, Z. The detectability of high-redshift Ly α emission lines prior to the reionization of the universe. *Astrophys. J.* **576**, L1–L4 (2002).
26. Miralda-Escudé, J., Haehnelt, M. & Rees, M. J. Reionization of the inhomogeneous universe. *Astrophys. J.* **530**, 1–16 (2000).
27. Loeb, A. & Rybicki, G. B. Scattered Lyman alpha radiation around sources before cosmological reionization. *Astrophys. J.* **524**, 527–535 (1999).
28. Haiman, Z. & Rees, M. J. Extended Lyman alpha emission around young quasars: A constraint on galaxy formation. *Astrophys. J.* **556**, 87–92 (2001).
29. Pentericci, L. *et al.* VLT optical and near-infrared observations of the $z = 6.28$ quasar SDSS J1030 + 0524. *Astron. J.* **123**, 2151–2158 (2002).
30. Brotherton, M. S., Wills, B. J., Steidel, C. C. & Sargent, W. L. W. Statistics of QSO broad emission-line profiles. 2: The C IV wavelength 1549, C III wavelength 1909, and MG II wavelength 2798 lines. *Astrophys. J.* **423**, 131–142 (1994).

Acknowledgements We thank E. Turner and H. Netzer for discussions, and are grateful for the hospitality of the Institute for Advanced Study where this work was completed. R.B. acknowledges the support of an Alon Fellowship at Tel Aviv University and of the Israel Science Foundation. A.L. acknowledges support from the Institute for Advanced Study and a John Simon Guggenheim Memorial Fellowship. This work was also supported by the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B. (e-mail: barkana@wise.tau.ac.il) or A.L. (e-mail: loeb@ias.edu).

Experimental extraction of an entangled photon pair from two identically decohered pairs

Takashi Yamamoto^{*†}, Masato Koashi^{*†}, Şahin Kaya Özdemir^{*†} & Nobuyuki Imoto^{*†‡}

^{*} School of Advanced Sciences, The Graduate University for Advanced Studies (SOKENDAI), Hayama, Kanagawa, 240-0193, Japan
[†] CREST Interacting Carrier Electronics Project, 4-1-8 Honmachi, Kawaguchi, 331-0012, Japan
[‡] NTT Basic Research Laboratories, NTT Corporation, 3-1 Morinosato-Wakamiya, Atsugi, 243-0198, Japan

Entanglement is considered to be one of the most important resources in quantum information processing schemes, including teleportation^{1–3}, dense coding⁴ and entanglement-based quantum key distribution⁵. Because entanglement cannot be generated by classical communication between distant parties, distribution of entangled particles between them is necessary. During the distribution process, entanglement between the particles is degraded by the decoherence and dissipation processes that result from unavoidable coupling with the environment. Entanglement distillation and concentration schemes^{6–9} are therefore needed to extract pairs with a higher degree of entanglement from these less-entangled pairs; this is accomplished using local operations and classical communication. Here we report an experimental demonstration of extraction of a polarization-entangled photon pair from two decohered

photon pairs. Two polarization-entangled photon pairs are generated by spontaneous parametric down-conversion and then distributed through a channel that induces identical phase fluctuations to both pairs; this ensures that no entanglement is available as long as each pair is manipulated individually. Then, through collective local operations and classical communication we extract from the two decohered pairs a photon pair that is observed to be polarization-entangled.

Several experimentally feasible schemes have been proposed for the extraction of photon pairs with a higher degree of entanglement from less-entangled pairs^{10–13} using local operations and classical communication (LOCC). Kwiat *et al.*¹⁴ have recently demonstrated that this is experimentally possible by local filtering. The local filtering is essentially an operation individually applied to each distributed pair: each pair goes through the filtering process and some of the pairs are discarded, while the surviving pairs have higher entanglement. The next step is to realize LOCC operations that are collectively applied to several pairs. The collective operations provide entanglement distillation and concentration for a broader class of noisy channels and with higher efficiency^{6–9}. Recently, we have proposed a feasible concentration scheme¹² based on a collective operation on two pairs with linear optics and photon detectors. In the following we will first briefly introduce this scheme and then describe the experimental realization.

Suppose that Alice and Bob want to share photon pairs in maximally entangled states:

$$|\Phi^{(\pm)}\rangle_{62} \equiv \frac{1}{\sqrt{2}}(|H\rangle_6|H\rangle_2 \pm |V\rangle_6|V\rangle_2) \quad (1)$$

where $|H\rangle$ and $|V\rangle$ represent horizontal and vertical polarization states, respectively. The subscript numbers represent the location of the photons in Fig. 1. Note that the states $|\Phi^{(+)}\rangle_{62}$ and $|\Phi^{(-)}\rangle_{62}$ can be easily converted to each other locally. Bob first prepares a maximally entangled photon pair in $|\Phi^{(+)}\rangle_{12}$, and sends its half to Alice through a quantum channel. This channel adds a phase shift ϕ that is fluctuating and hence is unknown. Bob then immediately prepares another pair in $|\Phi^{(+)}\rangle_{34}$ and sends its half through the same channel, which adds the same phase shift ϕ . When the phase shift is ϕ , the state after the transmission is:

$$|\phi\rangle_{12}|\phi\rangle_{34} \equiv \frac{1}{\sqrt{2}}(|H\rangle_1|H\rangle_2 + e^{i\phi}|V\rangle_1|V\rangle_2) \otimes \frac{1}{\sqrt{2}}(|H\rangle_3|H\rangle_4 + e^{i\phi}|V\rangle_3|V\rangle_4) \quad (2)$$

Because ϕ is unknown, the state actually shared by Alice and Bob is the following mixed state:

$$\rho_{1234} \equiv \int \frac{d\phi}{2\pi} |\phi\rangle_{12}\langle\phi| \otimes |\phi\rangle_{34}\langle\phi| \quad (3)$$

If we treat each pair separately, no entanglement is available. For example, the marginal state of the pair in modes 3 and 4 is given by:

$$\int \frac{d\phi}{2\pi} |\phi\rangle_{34}\langle\phi| = \frac{1}{2}(|HH\rangle_{34}\langle HH| + |VV\rangle_{34}\langle VV|) \quad (4)$$

where the off-diagonal elements are averaged out and the photons are obviously not entangled. On the other hand, if both pairs are considered together, there still remains an entanglement between Alice's two photons and Bob's two photons. In fact, it is easy to see that the whole state ρ_{1234} is a classical mixture of an entangled state $|HV\rangle_{13}|HV\rangle_{24} + |VH\rangle_{13}|VH\rangle_{24}$ and separable states $|HH\rangle_{13}|HH\rangle_{24}$ and $|VV\rangle_{13}|VV\rangle_{24}$. Here the relative phase between the terms $|HV\rangle_{13}|HV\rangle_{24}$ and $|VH\rangle_{13}|VH\rangle_{24}$ is not altered by the channel because both suffer the same amount of phase shift ϕ .

We can, in principle, extract the entangled component by LOCC in the following way. Alice first performs a collective projection measurement that gives the number of vertically polarized photons, without disclosing the polarization of each photon. If the outcome

of this measurement is unity, the post-measurement state is $|HV\rangle_{13}|HV\rangle_{24} + |VH\rangle_{13}|VH\rangle_{24}$, and then Alice and Bob can transform this state into an entangled photon pair in the form of equation (1) by applying local unitary operations. This scheme, called the Schmidt projection method, can be generally applied to n pairs of identically dephased photons or those identically prepared in a nonmaximally entangled pure state, with optimal efficiency in the asymptotic limit⁶.

Because it is difficult, using current technology, to distinguish the polarization of a photon without destroying it, the above scheme cannot be realized directly. We can, however, achieve almost the same function using a combination of linear optics and destructive photon detectors, shown in Fig. 1. Upon receiving the two photons, Alice first rotates the polarization of the photon in mode 3 by 90° using a half-wave plate (HWP₃). The photon in mode 1 passes through HWP₁, which simply adds a phase shift π (HWP₁ is inserted for a measurement of a single pair and is not relevant for the extraction experiment). At this point, the whole state is a mixture of $|HH\rangle_{13}|HV\rangle_{24} - |VV\rangle_{13}|VH\rangle_{24}$, $|HV\rangle_{13}|HH\rangle_{24}$, and $|VH\rangle_{13}|VV\rangle_{24}$. Alice then mixes the photons in modes 1 and 3 with a polarizing beam splitter (PBS) which transmits $|H\rangle$ and reflects $|V\rangle$. Note that if the state of the photons injected to PBS is $|HV\rangle_{13}$, mode 5 contains no photons, and if it is $|VH\rangle_{13}$, it contains two photons. On the other hand, if it is $|HH\rangle_{13}$ or $|VV\rangle_{13}$, it contains only one photon. Hence, by counting the number of photons in mode 5, we can discard the cases $|HV\rangle_{13}$ and $|VH\rangle_{13}$. To keep the entanglement in $|HH\rangle_{13}|HV\rangle_{24} - |VV\rangle_{13}|VH\rangle_{24}$, this measurement should not distinguish the polarization $|H\rangle$ from $|V\rangle$. Alice thus rotates the polarization by 45° at HWP₅ and detects photons on the diagonal basis $\{|D\rangle_5, |\bar{D}\rangle_5\}$, where $|D\rangle_5 \equiv (|H\rangle_5 + |V\rangle_5)/\sqrt{2}$ and $|\bar{D}\rangle_5 \equiv (|H\rangle_5 - |V\rangle_5)/\sqrt{2}$. Bob also rotates the polarization by 45° at HWP₄ and measures the polarization of the photon in mode 4 on the basis $\{|D\rangle_4, |\bar{D}\rangle_4\}$. Then, for example, if Alice finds only one photon in state $|D\rangle_5$ and Bob finds the photon in $|D\rangle_4$, modes 2 and 6 should contain a photon pair in the maximally entangled state $|\Phi^{(-)}\rangle_{62}$. In this way, Alice and Bob can extract entangled photon pairs by communicating the measurement results on modes 4 and 5.

In our experiments, photon pairs are generated using the schematics shown in Fig. 2. Two pairs of polarization-entangled photons

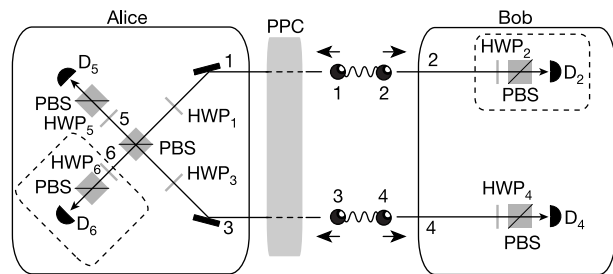


Figure 1 The schematic diagram of our experiment. Two pairs of polarization-entangled photons are distributed to Alice and Bob. Before reaching Alice, the photons in modes 1 and 3 pass through the pairwise phase-damping channel (PPC), which gives identical phase fluctuations to both photons. Alice manipulates the two photons collectively by using half-wave plates HWP₃ and HWP₅, a polarizing beam splitter (PBS), and a photon detector D₅. Bob detects a photon in mode 4 with D₄ after HWP₄ and PBS. HWP₃ rotates the polarization by 90°, while HWP₅ and HWP₄ rotate it by 45°. When each of the detectors D₄ and D₅ registers a photon, an entangled photon pair is emitted in modes 2 and 6. The apparatuses inside the dotted boxes are used for verification of the extracted entangled photon pair in modes 2 and 6. All detectors are silicon avalanche photodiodes, and they are placed after single-mode optical fibres to select a single spatial mode to ensure a high visibility. HWP₁ is inserted for an auxiliary experiment and it is not relevant here except for an additional phase shift π .

are generated by spontaneous parametric down-conversion from two adjacent Type I phase-matched β -barium borate (BBO) crystals^{15,16}. In this scheme, the desired entangled photon pairs (1&2 and 3&4) are generated whenever there is one photon at each output port of the non-polarizing beamsplitters. Note that this source works in a nondeterministic way, namely, there is a possibility of failure due to (1) no-photon generation in either or both of the BBOs, (2) channelling of both photons to one side of the beamsplitters, and (3) component losses. Such cases can be discarded by a post-selection, because at least one of modes 1, 2, 3 and 4 contains no photons.

After generating the two pairs of polarization-entangled photons, they are distributed to Alice and Bob. Before reaching Alice, the photons in modes 1 and 3 pass through the pairwise phase-damping channel (PPC), which gives identical phase fluctuations to both photons. The PPC is realized experimentally by inserting a liquid crystal retarder (LCR). It can provide any phase shift between the states $|H\rangle$ and $|V\rangle$ by changing the applied voltage. Eight different phases, which are equally spaced from 0 to 2π , are alternated every 10 s to simulate the phase-damping channel. Note that this realization of a PPC with an LCR is equivalent to a more practical situation in which modes 1 and 3 are the two sequential pulses travelling through the same optical fibre within the correlation time of the phase fluctuations.

In order to demonstrate that the PPC introduces the phase fluctuations and changes the marginal state of a single pair into the separable state of equation (4) when the LCR is modulated, we blocked the photons in modes 1 and 2 and measured the correlations between the polarizations of photons in modes 3 and 4 in Fig. 1 for an unmodulated and a modulated LCR. The results are shown in Fig. 3. We see that the $|H\rangle_3|H\rangle_4$ and $|V\rangle_3|V\rangle_4$ components are dominant. The coherence between these two terms should show up as an interference fringe in the count rate of D_4 when we rotate

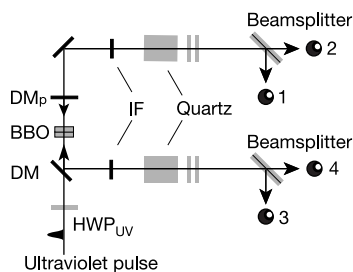


Figure 2 Experimental set-up for the polarization-entangled photon-pair source. Spontaneous parametric down-conversion from two adjacent Type I phase-matched, 2-mm-thick β -barium borate (BBO) crystals^{15,16} generates photon pairs collinearly. The ultraviolet light beam (average power 360 mW) used for pumping the BBO crystals for PDC is obtained from a frequency-doubled mode-locked Ti:sapphire laser (wavelength, 790 nm; pulse width, 80 fs; repetition rate, 82 MHz). The polarization of the pump beam is made diagonal to the axes of BBO using a half-wave plate (HWP_{UV}). After the first photon pair is generated, the pump beam is reflected back by a dichroic mirror (DM_p), and it passes the BBO in the opposite direction generating the second photon pair which is picked up by another dichroic mirror (DM). The group delay is compensated by thick quartz crystals, to erase the information on the origin (the first or the second BBO) of the photon pair¹⁶. The relative phase between H and V polarizations is adjusted by thin quartz crystals. Each photon pair is then split into two spatial modes by a non-polarizing beamsplitter to generate entangled photon pairs in modes 1&2 and 3&4. Temporal overlap of light beams from independent parametric down-conversions is a prerequisite for high-visibility interference^{22–24}. In our scheme, this is achieved by spectral filtering using narrow-band interference filters (IF; wavelength, 790 nm; bandwidth, 3.5 nm) as it has been done in other multiphoton interference experiments^{2,18–21}. The timing between the photons in modes 1 and 3 is precisely adjusted by mounting the DM_p on a motorized stage.

HWP_4 , on condition that a photon in state $|D\rangle_3$ (or $|\bar{D}\rangle_3$) is detected in mode 3. The result shows a visibility of 0.89 when the LCR is unmodulated, implying the creation of highly entangled photon pairs. When the LCR is modulated, visibility becomes less than 0.03, which is a good sign of the loss of coherence. A measurement on the photon pair in modes 1 and 2 has produced similar results. These results are further supported by the tomographic reconstruction of the density matrix¹⁷ as shown in Fig. 3c for the modulated LCR. From the reconstructed density matrix for this decohered pair, we can calculate the entanglement of formation (EOF). EOF is a measure of entanglement taking the value of unity for a maximally entangled qubit pair and zero for separable states. For our decohered photon pair, EOF is calculated to be very small; $EOF = 0.0018$. This clearly indicates that the photon pairs shared by Alice and Bob in our experiment do not allow the extraction of a highly entangled pair as long as each pair is manipulated individually.

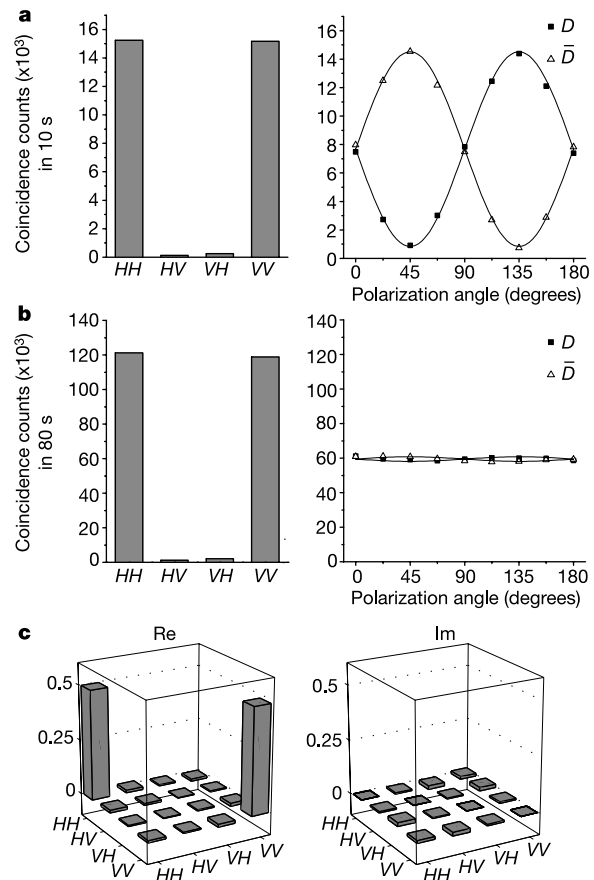


Figure 3 Experimental results showing that PPC decoheres individual pairs. The modes 1 and 2 are blocked, and the photons in modes 3 and 4 are measured after passing through the PPC. The axis of HWP_5 is adjusted so that it does not rotate the polarization. Then polarization correlations are recorded by coincidence counting between detectors D_5 and D_4 for various angles of HWP_3 and HWP_4 . **a, b**, Typical coincidence measurements for unmodulated and modulated LCR, respectively. Left, coincidence rates on the $\{|H\rangle, |V\rangle\}$ basis. Right, the rate of coincidence events where a photon in polarization $|D\rangle$ (or $|\bar{D}\rangle$) is detected at D_5 and a linearly polarized photon with angle θ is detected at D_4 ($\theta = 0^\circ$, $\theta = 45^\circ$, $\theta = 90^\circ$ and $\theta = 135^\circ$ correspond to $|H\rangle$, $|D\rangle$, $|V\rangle$ and $|\bar{D}\rangle$, respectively). The solid curve is the best fit to the data by two sine functions with the same amplitude. The observed visibility is 0.89 in **a** and less than 0.03 in **b**. **c**, Real and imaginary components of the density matrix. Reconstruction is done by recording polarization correlations measured on 16 different settings of HWP_3 , HWP_4 and additionally inserted quarter-wave plates in modes 3 and 4. This density matrix is close to the state in equation (4). The entanglement of formation (EOF) calculated from the density matrix is 0.0018.

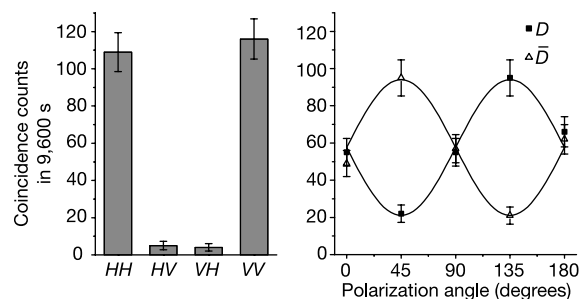


Figure 4 Experimental results showing that the extracted photon pair is entangled. Polarization correlations between the photon pair in modes 2 and 6 are taken on condition that the detectors D_4 and D_5 both register a photon. The left shows coincidence rates on the $\{|H\rangle, |V\rangle\}$ basis. The right is the rate of coincidence events where a photon in polarization $|D\rangle$ (or $|\bar{D}\rangle$) is detected at D_6 and a linearly polarized photon with angle θ is

detected at D_2 . The error bars assume the Poisson statistics of the events. The solid curve represents the best fit to the data. The observed visibility 0.63 ± 0.05 demonstrates that the photons in modes 2 and 6 are entangled. The lower bound of fidelity to the state $|\Phi^{(-)}\rangle_{62}$ is 0.78 ± 0.05 , from which the lower bound of EOF is calculated as 0.42 ± 0.12 .

Now we will show that the collective manipulation shown in Fig. 1 can extract an entangled photon pair in modes 2 and 6 by selecting the cases where a photon in state $|D\rangle_5$ is observed in mode 5 and a photon in $|D\rangle_4$ is observed in mode 4. Because our source of entangled pairs based on parametric down-conversion in Fig. 2 is nondeterministic, we should post-select the events where both of the photon detectors at modes 2 and 6 register a photon. Hence, what we have measured is a fourfold coincidence at detectors D_2 , D_4 , D_5 and D_6 . In this case, there is no need to distinguish the arrival of two photons in mode 5, because such events (no photons in mode 6) are discarded by the post-selection. The results of the polarization correlations for the selected photon pair in modes 2 and 6 are shown in Fig. 4. In contrast to Fig. 3b, it clearly shows an interference fringe. Visibility is found to be 0.63 ± 0.05 , which is well above the achievable value of 0.5 in the classical model. We believe that the visibility is limited mainly by the residual temporal and spatial mode mismatch between the photons belonging to different pairs, as is the case in other interference experiments^{2,18–21} and theoretical studies^{22–24}.

The rate of the fourfold coincidence rate is quite low, so we have not collected enough data sets to reconstruct the density matrix of the extracted pair. Nevertheless, we can still set a bound²⁵ to the fidelity f of the extracted pair to the state $|\Phi^{(-)}\rangle_{62}$ as $0.78 \pm 0.05 \leq f \leq 0.89 \pm 0.02$, from the observed correlations on $\{|H\rangle, |V\rangle\}$ and $\{|D\rangle, |\bar{D}\rangle\}$ bases. The lower bound of the fidelity can then be used to determine a lower bound for EOF⁷, which turns out to be $\text{EOF} \geq 0.42 \pm 0.12$. The extracted photon pair in our experiment is thus clearly shown to be entangled.

The collective operation on two photons used in our experiment to extract entanglement is essentially the one referred to as quantum parity check^{26,27}. This feature was originally proposed in ref. 28, and forms an important building block for quantum information processing by linear optics^{11,29,30}. Our experiment has shown that the quantum parity check can be applied to a nontrivial task, and we hope that this demonstration forms a step towards experimental realizations of more complicated applications in quantum communication and computation. □

Received 25 October; accepted 29 November 2002; doi:10.1038/nature01358.

1. Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1898 (1993).
2. Bouwmeester, D. *et al.* Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
3. Furusawa, A. *et al.* Unconditional quantum teleportation. *Science* **282**, 706–709 (1998).
4. Bennett, C. H. & Wiesner, S. J. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Phys. Rev. Lett.* **69**, 2881–2884 (1992).
5. Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).
6. Bennett, C. H., Bernstein, H. J., Popescu, S. & Schumacher, B. Concentrating partial entanglement by local operations. *Phys. Rev. A* **53**, 2046–2052 (1996).

7. Bennett, C. H., DiVincenzo, D. P., Smolin, J. A. & Wootters, W. K. Mixed-state entanglement and quantum error correction. *Phys. Rev. A* **54**, 3824–3851 (1996).
8. Bennett, C. H. *et al.* Purification of noisy entanglement and faithful teleportation via noisy channels. *Phys. Rev. Lett.* **76**, 722–725 (1996).
9. Deutsch, D. *et al.* Quantum privacy amplification and the security of quantum cryptography over noisy channels. *Phys. Rev. Lett.* **77**, 2818–2821 (1996).
10. Bose, S., Vedral, V. & Knight, P. L. Purification via entanglement swapping and conserved entanglement. *Phys. Rev. A* **60**, 194–197 (1999).
11. Pan, J.-W., Simon, C., Brukner, C. & Zeilinger, A. Entanglement purification for quantum communication. *Nature* **410**, 1067–1070 (2001).
12. Yamamoto, T., Koashi, M. & Imoto, N. Concentration and purification scheme for two partially entangled photon pairs. *Phys. Rev. A* **64**, 012304 (2001).
13. Zhao, Z., Pan, J.-W. & Zhan, M. S. Practical scheme for entanglement concentration. *Phys. Rev. A* **64**, 014301 (2001).
14. Kwiat, P. G., Barraza-Lopez, S., Stefanov, A. & Gisin, N. Experimental entanglement distillation and 'hidden' non-locality. *Nature* **409**, 1014–1017 (2001).
15. Kwiat, P. G. *et al.* Ultrabright source of polarization-entangled photons. *Phys. Rev. A* **60**, R773–R776 (1999).
16. Kim, Y. H., Kulik, S. P. & Shih, Y. High-intensity pulsed source of space-time and polarization double-entangled photon pairs. *Phys. Rev. A* **62**, 011802 (2000).
17. James, D. F. V., Kwiat, P. G., Munro, W. J. & White, A. G. Measurement of qubits. *Phys. Rev. A* **64**, 052312 (2001).
18. Bouwmeester, D. *et al.* Observation of three-photon Greenberger-Horne-Zeilinger entanglement. *Phys. Rev. Lett.* **82**, 1345–1349 (1999).
19. Pan, J.-W. *et al.* Experimental demonstration of four-photon entanglement and high-fidelity teleportation. *Phys. Rev. Lett.* **86**, 4435–4438 (2001).
20. Jennewein, T., Weihs, G., Pan, J.-W. & Zeilinger, A. Experimental nonlocality proof of quantum teleportation and entanglement swapping. *Phys. Rev. Lett.* **88**, 017903 (2002).
21. Lamas-Linares, A., Howell, J. C. & Bouwmeester, D. Stimulated emission of polarization-entangled photons. *Nature* **412**, 887–890 (2001).
22. Zukowski, M., Zeilinger, A. & Weinfurter, H. Entangling photons radiated by independent pulsed sources. *Ann. NY Acad. Sci.* **755**, 91–102 (1995).
23. Rarity, J. G. Interference of single photons from separate sources. *Ann. NY Acad. Sci.* **755**, 624–631 (1995).
24. Özdemir, Ş. K., Miranowicz, A., Koashi, M. & Imoto, N. Pulse mode projection synthesis: Effects of mode mismatch on optical state truncation and preparation. *Phys. Rev. A* **66**, 053809 (2002).
25. Nagata, K., Koashi, M. & Imoto, N. Observables suitable for restricting the fidelity to multipartite maximally entangled states. *Phys. Rev. A* **65**, 042314 (2002).
26. Pittman, T. B., Jacobs, B. C. & Franson, J. D. Probabilistic quantum logic operations using polarizing beam splitters. *Phys. Rev. A* **64**, 062311 (2001).
27. Pittman, T. B., Jacobs, B. C. & Franson, J. D. Demonstration of nondeterministic quantum logic operations using linear optical elements. *Phys. Rev. Lett.* **88**, 257902 (2002).
28. Pan, J.-W. & Zeilinger, A. Greenberger-Horne-Zeilinger-state analyzer. *Phys. Rev. A* **57**, 2208–2211 (1998).
29. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
30. Koashi, M., Yamamoto, T. & Imoto, N. Probabilistic manipulation of entangled photons. *Phys. Rev. A* **63**, 030301 (2001).

Acknowledgements We thank K. Nagata, K. Tamaki, A. Miranowicz and J. Shimamura for helpful discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.I. (e-mail: imoto@soken.ac.jp).

Direct observation of a local thermal vibration anomaly in a quasicrystal

Eiji Abe^{*†}, S. J. Pennycook^{*} & A. P. Tsai[†]

^{*} Condensed Matter Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA

[†] National Institute for Materials Science and SORST, Japan Science and Technology Corporation, 1-2-1 Sengen, Tsukuba 305-0047, Japan

Quasicrystals have long-range order with symmetries that are incompatible with periodicity, and are often described with reference to a higher-dimensional analogue of a periodic lattice^{1–3}. Within the context of this ‘hyperspace’ crystallography, lattice dynamics of quasicrystals can be described by a combination of lattice vibrations and atomic fluctuations—phonons and phasons^{1,4}. However, it is difficult to see localized fluctuations in a real-space quasicrystal structure, and so the nature of phason-related fluctuations and their contribution to thermodynamic stability are still not fully understood. Here we use atomic-resolution annular dark-field scanning transmission electron microscopy to map directly the change in thermal diffuse scattering intensity distribution in the quasicrystal, through *in situ* high-temperature observation of decagonal Al₇₂Ni₂₀Co₈. We find that, at 1,100 K, a local anomaly of atomic vibrations becomes significant at specific atomic sites in the structure. The distribution of these localized vibrations is not random but well-correlated, with a quasiperiodic length scale of 2 nm. We are able to explain this feature by an anomalous temperature (Debye–Waller) factor for the Al atoms that sit at the phason-related sites defined within the framework of hyperspace crystallography. The present results therefore provide a direct observation of local thermal vibration anomalies in a solid.

Within the hyperspace n -dimensional ($n \geq 4$) crystallography, the quasiperiodic structure can be generated from the n -dimensional periodic lattice decorated by so-called occupation domains⁵ (or atomic surfaces) that can be converted into the atomic arrangement of the real quasicrystal. The n -dimensional crystal is defined by n basis vectors—the number of vectors is more than the real dimensions so as to incorporate the quasicrystal symmetry, giving rise to a unique elastic degree of freedom specific to quasicrystals; this results in ‘phasons’, in addition to the usual phonons that are present in crystals. The phason has no counterpart in periodic crystals, and hence the Debye–Waller (DW) factor of a quasicrystal may have both phonon and phason contributions^{1,5,6}, written as;

$$M = M_{\text{phonon}} + M_{\text{phason}} \quad (1)$$

Here, M is the DW factor defined by mean-square thermal vibration amplitude, $\langle u^2 \rangle$, of the atoms. The occupation domains are generally not continuous in quasicrystals, and a phason model predicts a significant fluctuation that smears the surfaces (edges) of the occupation domains. This surface fluctuation causes a local anomaly of the DW factor at the corresponding real-space atomic sites, but does not destroy the long-range quasiperiodic order; it causes a reduction in the Bragg intensity, and consequently some additional diffuse scattering.

The existence of phason-related atomic fluctuations is a critical factor in the thermodynamic stability of quasicrystals, whatever their origin; the context could be a random-tiling-like realization⁷ where phason fluctuations are essential to provide entropic stabilization, or an unlocked state⁸ of an intrinsically energy-minimized perfect quasicrystal. In both cases, phason fluctuations are expected to be significant above some critical temperature. So far, numerous experimental measurements of DW factors^{9,10}, phason atomic jumps^{11,12}, large-scale (~ 2 nm) phason tile-flips¹³ and phason-

related structural defects^{14,15} have been attempted by various techniques. However, there has been no direct evidence on where such localized atomic fluctuations take place in the real-space quasicrystal structure.

Annular dark-field scanning transmission electron microscopy (ADF-STEM) provides atomic-resolution images by effectively illuminating each atomic column one-by-one as a finely focused probe is scanned across a thin crystalline specimen. The annular detector generates an intensity map of the scattered electrons with atomic-number (Z) contrast¹⁶. Because the detector is at a high angle, the scattering is dominated by phonon scattering events (that is, thermal diffuse scattering, TDS), and is therefore very sensitive to the DW factor; an Einstein model of independently vibrating atoms is valid for describing the multiphonon contribution that dominates high-angle diffuse scattering. Thus, the TDS cross-section (σ_{TDS}) over the range of the annular detector is given by¹⁶:

$$f'_{\text{HA}}(M, s) \approx \sigma_{\text{TDS}} \propto \int_{\text{detector}} f^2(s) [1 - \exp(-2Ms^2)] d^2s \quad (2)$$

where $f'_{\text{HA}}(M, s)$ is the absorptive atomic form factor for high-angle scatterings, $f(s)$ is the atomic form factor for elastic scatterings (with $s = \theta/2\lambda$, where θ is the scattering angle and λ is the electron wavelength) and M is the DW factor. The thermal vibration is responsible for destroying the coherence along the beam direction, and hence the scattering is assumed to be generated incoherently from individual atoms. Thus, to a good approximation, the intensity of each illuminated atomic column will be directly dependent on the σ_{TDS} within the relevant column (although TDS described by σ_{TDS} does not reflect any fine details¹⁷ of the high-angle diffraction pattern, it is sufficient for estimating the integrated intensity reaching the detector).

Here we present the direct observation of local anomalies of the DW factor in a quasicrystal, through an *in situ* high-temperature ADF-STEM observation of decagonal Al₇₂Ni₂₀Co₈ for which a nearly perfect, ideal quasiperiodic structure (referred to as a basic AlNiCo structure) can be obtained by annealing at 1,100 K (the thermodynamically stable phase)¹⁸. Because the decagonal structure has a two-dimensional character described as a periodic stacking of quasiperiodic planes, ADF-STEM with the incident beam along a ten-fold symmetry axis enables us to observe directly the quasiperiodic arrangement of atoms^{19,20}. In the observation at room temperature (300 K) of the water-quenched sample after annealing at 1,100 K (Fig. 1a), the ADF-STEM image highlights the transition metal (TM: Ni or Co) positions relative to the Al, owing to the $f^2(s)$

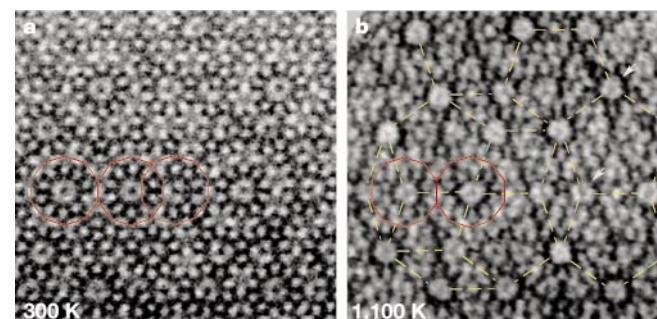


Figure 1 Atomic-resolution annular dark-field scanning transmission electron microscope (ADF-STEM) images of decagonal Al₇₂Ni₂₀Co₈. These were taken at **a**, 300 K and **b**, 1,100 K, by collecting the electrons scattered at angles between approximately 45 and 100 mrad ($0.9 \leq s \leq 2.0$) with a 200-kV STEM. The images were obtained, although not from exactly the same region, from the near-edge regions of the same cleavage grain, so that the specimen thickness is comparable. Therefore, contrast differences between **a** and **b** are due to the different temperature, not to changes in specimen thickness. By connecting the centre of 2-nm decagonal clusters (red) that reveal significant temperature-dependent contrast change, pentagonal quasiperiodic tiling (yellow) with an edge length of 2 nm can be drawn in **b**.

dependence of the contrast (equation (2)). Tracing the major brightest spots provided an excellent match to a perfect quasiperiodic pattern described by overlapping decagonal tiles²¹ (Gummelt tiling²²) with a diameter of 2 nm, some of which are outlined in Fig. 1a, leading to a quasi-unit-cell model structure based on the atomic cluster with broken ten-fold symmetry^{21,23,24}. But when the sample is heated and held at a temperature of approximately 1,100 K in the microscope, we find a remarkable change in the relative contrast (compare Fig. 1b with Fig. 1a). It is evident that a significant enhancement in contrast appears at some specified places that can be well represented by the pentagonal Penrose tiling with an edge length of 2 nm, although a local matching rule seems to be not completely maintained (note that pentagonal Penrose tiling is uniquely related to Gummelt's tiling by connecting the specified decagons separated by 2 nm).

Further, we notice that anomalous contrast occurs at cores of the decagons (Fig. 1b); a representative feature is shown in Fig. 2b and d. (A few other clusters reveal complicated contrasts at the core, including five-fold symmetry-like appearance of the bright spots indicated by arrows in Fig. 1b, although these clusters are rare compared to the clusters shown in Fig. 2. Possible interpretations of this issue will be discussed later.) Viewing carefully the interior contrast of the decagon in Fig. 2b and d, an increase of intensity is found to be significant at the positions indicated by arrowheads, which are the Al sites in the decagonal cluster model²³ (Fig. 2e). The intensity profiles show that the Al atom at the core, denoted as Al α ,

in fact shows stronger contrast at 1,100 K than that at 300 K. We confirmed that this temperature-dependent contrast change is reversible; that is, after cooling down to 300 K from 1,100 K, the anomalous contrast regions become darker again. It should be noted that the intensity at the Al α site ($I_{\text{Al}\alpha}$) is originally stronger than that of the other Al sites at 300 K (Fig. 2b); the $I_{\text{Al}\alpha}$ is found to vary significantly depending on the angular range of the detector (Fig. 2a–c), whereas the intensities of the other Al and TM atoms do not— $I_{\text{TM}}/I_{\text{Al}}$ is almost constant at any angle-range. These angle-dependent as well as temperature-dependent (reversible) anomalous contrasts can naturally be attributed to a local anomaly of the DW factor, as expected through equation (2); assuming that the DW factor effect is equivalent for all the Al sites, neither of these dependences is expected.

We now interpret the significant increase of $I_{\text{Al}\alpha}$ in terms of atomic vibration amplitudes, $\langle u^2 \rangle$, for the observations both at 300 K and 1,100 K; we assume that there is no static column distortion effect¹⁶ contributing to the anomalous $I_{\text{Al}\alpha}$ even at 300 K. (Static transverse displacements of the Al α atoms cause relaxations of the nearest-neighbour TM atoms; these TM atoms are also distorted, and are expected to reveal anomalous contrast compared to the other TM atoms; however, they do not show any significant angle-dependent contrast change— $I_{\text{TM}}/I_{\text{Al}}$ is almost constant in Fig. 2a–c.) Note that $I_{\text{TM}}/I_{\text{Al}}$ does not change much, even after increasing the temperature (Fig. 2d), indicating that the effects of temperature on the DW factors of these Al and TM atoms

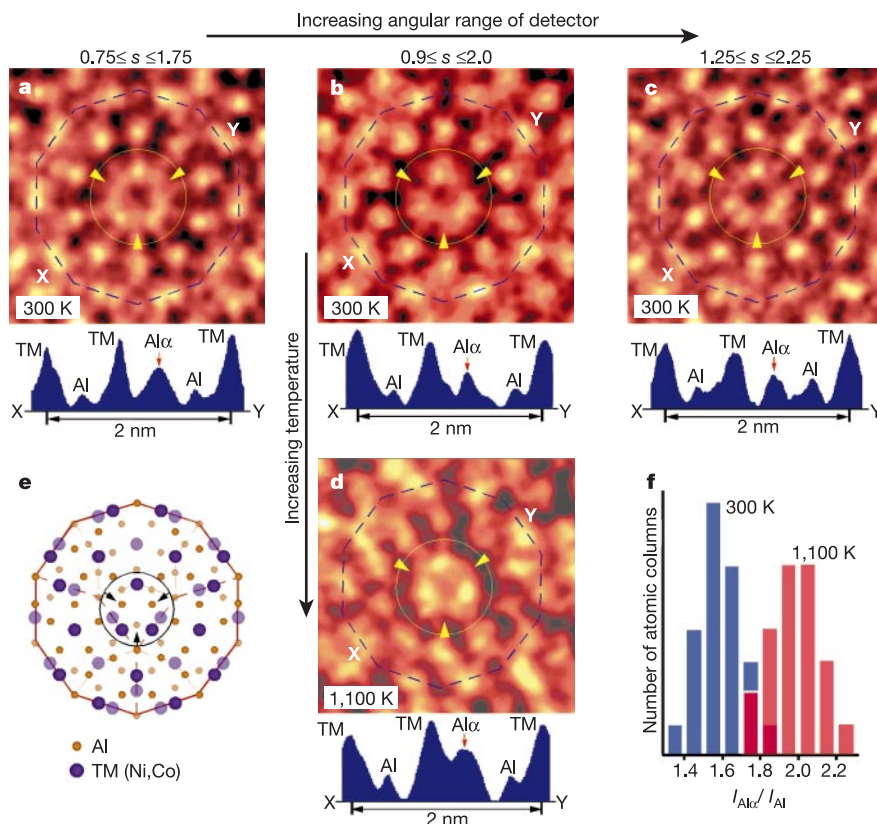


Figure 2 A decagonal cluster with a diameter of about 2 nm, a structural unit of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$. The clusters were observed at 300 K (**a–c**) with different angular ranges of the detector, and at 1,100 K (**d**); also shown are the corresponding line profiles across X–Y. Images of **a** and **c** are obtained from the same cluster, using the 300-kV STEM with detector angle ranges of about 30–70 mrad ($0.75 \leq s \leq 1.75$) and 50–90 mrad ($1.25 \leq s \leq 2.25$), respectively. Images **b** and **d** are enlargements of a part of Fig. 1a and b, respectively. The images **a–d** show significant contrast changes at the core of the cluster, particularly at the positions indicated by arrows, which correspond to the Al α sites denoted in the line profiles. **e**, Structural model of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ with a practical averaged configuration derived by superimposing all possible atomic positions in the quasi-unit-cell

model (assuming perfect quasiperiodic atomic order, there appeared to be three variations of atomic arrangements in the interior of the cluster²³). Structural refinements^{27,28} based on single-crystal X-ray data converge on a nearly identical arrangement to this averaged configuration. The structure has two distinct atomic layers stacked along the periodic *c*-axis; bold and pale dots denote the atoms at layers $c = 0$ and $c = 1/2$, respectively. **f**, Histograms of intensity ratios between the Al α and Al atomic columns, $I_{\text{Al}\alpha}/I_{\text{Al}}$, from the 200-kV STEM images at 300 K and 1,100 K. The column intensities were measured for about 30 decagonal clusters, revealing similar behaviour to that seen in **b** and **d**; these features are typical of the clusters located at the 2-nm-scale pentagonal Penrose lattice.

are almost equivalent, and small compared with that of the $\text{Al}\alpha$. Thus, we normalize the $I_{\text{Al}\alpha}$ with reference to the I_{Al} at each temperature. In Fig. 2f we show histograms of $I_{\text{Al}\alpha}/I_{\text{Al}}$ distributions, in which the $I_{\text{Al}\alpha}/I_{\text{Al}}$ peaks at around ~ 1.5 and ~ 2.0 (assuming a gaussian distribution) for the observations at 300 K and 1,100 K, respectively. Both the temperature and angle dependences are fairly well explained by the σ_{TDS} for different $\langle u^2 \rangle$ values; see the $\sigma_{\text{TDS}}^{\text{Al}} - \langle u^2 \rangle$ curves calculated on the basis of equation (2) (Fig. 3).

With the angle range of $0.9 \leq s \leq 2.0$, the ratio of $\sigma_{\text{TDS}}^{\text{Al}}$ between the standard Al (denoted as Als whose DW factor is assumed not to be temperature dependent) and the $\text{Al}\alpha$, $\sigma_{\text{TDS}}^{\text{Al}\alpha}/\sigma_{\text{TDS}}^{\text{Al}}$, can be approximately 1.5 and 2.0 with the appropriate $\langle u^2 \rangle$ values at each temperature, for example, $\sim 0.4 \times 10^{-4} \text{ nm}^2$ (Als), $\sim 0.9 \times 10^{-4} \text{ nm}^2$ ($\text{Al}\alpha$ at 300 K) and $> \sim 3.0 \times 10^{-4} \text{ nm}^2$ ($\text{Al}\alpha$ at 1,100 K); this reproduces well the observed temperature-dependent change of $I_{\text{Al}\alpha}/I_{\text{Al}}$ (note that these $\langle u^2 \rangle$ values are only tentative and may be improved when a more accurate $f'_{\text{HA}}(M, s)$ is available). A systematic angle-dependent change of $I_{\text{Al}\alpha}/I_{\text{Al}}$ observed at 300 K is also explained well by these $\langle u^2 \rangle$ differences; the $\text{Al}\alpha$ can reveal stronger contrast, relative to that of the Als, for the lower angle ranges (see red arrows in Fig. 3). Although the $\langle u^2 \rangle$ at 1,100 K cannot be determined because σ_{TDS} saturates at large $\langle u^2 \rangle$ values, we nevertheless emphasize that the observed anomalous contrast at the $\text{Al}\alpha$ site has been clearly correlated with differences in $\langle u^2 \rangle$. This is, to our knowledge, the first direct observation of a local vibration anomaly in a solid.

Significantly large $\langle u^2 \rangle$ at high temperature is consistent with the recent X-ray diffuse scattering measurements on the same sample of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$, in which the overall $\langle u^2 \rangle$ is suggested to be of the order of $\sim 1.0 \times 10^{-3} \text{ nm}^2$ (at 1,100 K) when compared with a uniform harmonic vibration²⁵. The local anomalies of $\langle u^2 \rangle$ in the present work imply significant anharmonicity at the $\text{Al}\alpha$ site. Similar anomalies might therefore be anticipated at the neighbouring TM sites, and some evidence of this is seen at 1,100 K (Fig. 2d). Further supporting evidence comes from recent molecular-dynamics simulations²⁶ of the structure at 1,000 K, which predicted the occurrence of DW factor anomalies for the Al atoms at the core of the 2-nm decagonal cluster.

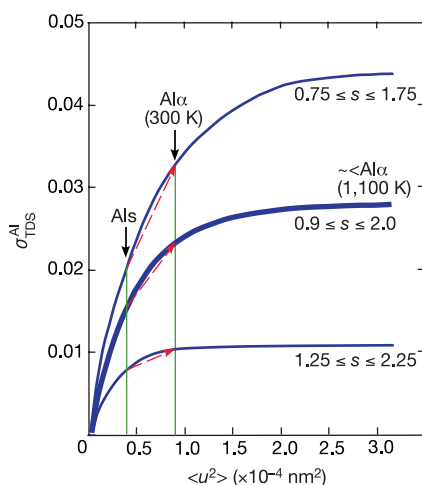


Figure 3 Estimates of the thermal diffuse scattering cross-section of Al ($\sigma_{\text{TDS}}^{\text{Al}}$; equation (2)) as a function of mean-square thermal vibration amplitude ($\langle u^2 \rangle$). Data are shown for several detector angle (s) ranges (blue curves), and were calculated using the atomic form factor, $f(s)$, in ref. 30, by assuming an isotropic Debye–Waller (DW) factor ($M = 8\pi^2 \langle u^2 \rangle$). The bold curve ($0.9 \leq s \leq 2.0$) explains the temperature-dependent contrast change by assuming the appropriate $\langle u^2 \rangle$ values for the Als (green line), $\text{Al}\alpha$ at 300 K (green line) and $\text{Al}\alpha$ at 1,100 K ($> \sim 3.0 \times 10^{-4} \text{ nm}^2$). Angle-dependent changes (at 300 K) expected from these $\langle u^2 \rangle$ values are indicated by red arrows. Details are described in the text.

The $\text{Al}\alpha$ atoms, located at the centre of the decagonal clusters that are on the 2-nm-scale pentagonal quasiperiodic lattice vertices, are generated from the edge of the occupation domains (α in Fig. 4) within the hyperspace crystallographic description²⁷. Thus, the present local DW factor anomaly can be attributed to significant fluctuations of the occupation domains in specified surface regions, which, in turn, indicate an occurrence of phasonic fluctuations—a perturbation of quasiperiodic order that can be described through M_{phason} (equation (1))—realized for the $\text{Al}\alpha$ atoms at the centre of the 2-nm cluster. Some Al atoms within the 2-nm cluster are also placed in a similar local environment to that of the $\text{Al}\alpha$ atoms—see the ‘kite’²³ tiles drawn in Fig. 2e. However, they are symmetrically (crystallographically) not equivalent to the $\text{Al}\alpha$ site, and hence may not reveal any significant anomalies (that is, the atomic configuration in the kite tiles can be different²³ depending on their local-neighbour environment). An occurrence of DW factor anomalies at the $\text{Al}\alpha$ site is probably induced by the presence of the phason-flip atomic sites, denoted as β in Fig. 4, which are separated by less than a typical interatomic distance. Therefore the α and β sites cannot be occupied simultaneously, and the β sites, considered to be energetically similar to the $\text{Al}\alpha$ site²⁶, could act as vacancies in providing an effective space for relaxation; it is reasonable to assume that this causes significant anisotropy in the DW factor (Fig. 4). Although the resolution of the present STEM ($\sim 1.5 \text{ \AA}$) is not sufficient to reveal this anisotropic behaviour, molecular-dynamics structural simulations²⁶ at 1,000 K have indicated an anisotropic DW factor

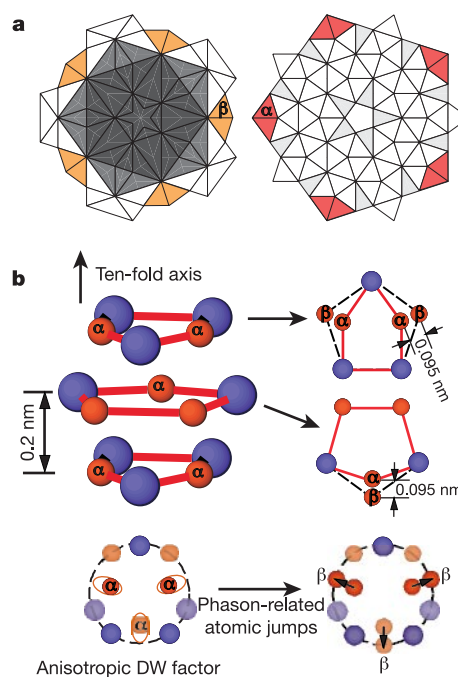


Figure 4 ‘Phason’-related atomic sites in the $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ quasicrystal. **a**, Higher-dimensional, or ‘hyperspace’, crystallographic description of the decagonal $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ structure²⁷. Occupation domains are placed at $(1/5, 1/5, 1/5, 1/5, 1/4)$ (left) and $(2/5, 2/5, 2/5, 2/5, 1/4)$ (right) in the five-dimensional decagonal lattice with successful partitions assigned by Al and TM (transition metal), where the grey level represents the concentration of the TM atoms at relevant sites (see ref. 27 for details). Extra β portions (yellow) have been added to the edges of the occupation domain to generate phason-flip atomic sites with respect to the α sites that are generated from the α portions (red). **b**, Pentagonal columnar atomic configuration around the centre of the 2-nm decagonal cluster; orange and purple atoms represent Al and TM, respectively. Phason-related α and β sites are on the quasiperiodic atomic plane, separated by approximately 0.095 nm. Possible anisotropy of the DW factor of the Al atoms at the α sites and occurrence of atomic jumps into the β sites are described in the projected atomic positions shown below, where the pale atoms represent those at different levels along the ten-fold axis.

identical to that shown in Fig. 4.

Regarding diffusional atomic jumps at high temperature, it is natural to consider that the local vibration anomalies enhance the short-range diffusion of atoms located around the cluster centre with a help of a phason-site-mediated transport mechanism, causing occasional diffusion jumps of not only Al but also TM atoms into the phason β site. This explains a random occurrence of five-fold symmetry-like clusters at some places (Fig. 1b). Note that the atoms delivered to the β site by diffusion jump can be quenched in, so that they will be recognized as 'quenched phason-defects' (including chemical and site-occupation disorders that are significant at the core of the 2-nm clusters), which have in fact been detected by experimental measurements on the quenched sample^{20,27,28}.

Alternatively, five-fold-symmetric 2-nm clusters emerge as the thermodynamically stable configuration when the present high-temperature $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ undergoes phase transition at lower temperatures (≤ 990 K, to a less-quasiperiodic ordered phase); this suggests that the phason β sites, metastable at high temperature, provide stable positions for TM atoms at low temperature. This also indicates that the above structural transition is closely related to phason-related atomic behaviour, which is significant at the centre of the 2-nm clusters, as shown in the present direct observation. □

Methods

Atomic-resolution images were taken by 200-kV (JEOL 2010F) and 300-kV (VG Microscopes, HB603U) scanning transmission electron microscopes, providing a minimum probe of approximately 0.150 nm and 0.126 nm, respectively. The former microscope was used for a high-temperature *in situ* observation. Specimens were prepared by crushing the bulk material and depositing it onto perforated amorphous carbon film supported on Cu grids; by this method, the surface amorphous layer, roughness, and contamination that are frequently induced by ion-milling and which strongly affect image contrast, can be avoided. Thus, the intensities measured from the present images give reliable information. Heating to 1,100 K in the microscope was achieved within 5 min, which effectively suppresses a phase transition that would occur at ≤ 990 K during heating. The electron diffraction pattern taken at 1,100 K from the grain used for imaging (Fig. 1) confirmed that $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ retains its basic AlNiCo structure during *in situ* high-temperature observation. Energy-dispersive X-ray spectroscopy showed no detectable compositional change of the grain before and after the heat treatment in the microscope.

Received 20 September; accepted 25 November 2002; doi:10.1038/nature01337.

1. Bak, P. Phenomenological theory of icosahedral incommensurate ("quasiperiodic") order in Mn-Al alloys. *Phys. Rev. Lett.* **54**, 1517–1519 (1985).
2. Janssen, T. Crystallography of quasi-crystals. *Acta Crystallogr. A* **42**, 261–271 (1986).
3. Yamamoto, A. Crystallography of quasiperiodic crystals. *Acta Crystallogr. A* **52**, 509–560 (1996).
4. Soclar, T., Lubensky, T. & Steinhardt, P. J. Phonons, phasons, and dislocations in quasicrystals. *Phys. Rev. B* **34**, 3345–3360 (1986).
5. Jaric, M. V. & Nelson, D. R. Diffuse scattering from quasicrystals. *Phys. Rev. B* **37**, 4458–4472 (1988).
6. Ishii, Y. Phason softening and structural transitions in icosahedral quasicrystals. *Phys. Rev. B* **45**, 5228–5239 (1992).
7. Henley, C. L. in *Quasicrystals: The State of the Art* (eds DiVincenzo, D. & Steinhardt, P. J.) 429–524 (World Scientific, Singapore, 1991).
8. Jeong, H. C. & Steinhardt, P. J. Finite-temperature elasticity phase transition in decagonal quasicrystals. *Phys. Rev. B* **48**, 9394–9403 (1993).
9. Bancel, P. A. in *Quasicrystals: The State of the Art* (eds DiVincenzo, D. & Steinhardt, P. J.) 17–55 (World Scientific, Singapore, 1991).
10. Colella, R., Zhang, Y., Sutter, J. P., Ehrlich, S. N. & Kycia, S. W. Debye-Waller factors in a quasicrystal. *Phys. Rev. B* **63**, 014202 (2000).
11. Dolinske, J., Apih, T., Simsic, M. & Dubois, J. M. Self-diffusion in icosahedral $\text{Al}_{72.4}\text{Pd}_{20.5}\text{Mn}_{7.1}$ and phason percolation at low temperatures studied by ^{27}Al NMR. *Phys. Rev. Lett.* **82**, 572–575 (1999).
12. Coddens, G. & Steurer, W. Time-of-flight neutron-scattering study of phason hopping in decagonal Al-Co-Ni quasicrystals. *Phys. Rev. B* **60**, 270–276 (1999).
13. Edagawa, K., Suzuki, K. & Takeuchi, S. High resolution transmission electron microscopy observation of thermally fluctuating phasons in decagonal Al-Cu-Co. *Phys. Rev. Lett.* **85**, 1674–1677 (2000).
14. de Boissieu, M. *et al.* Diffuse scattering and phason elasticity in the AlPdMn icosahedral phase. *Phys. Rev. Lett.* **75**, 89–92 (1995).
15. Zeger, G., Plachke, D., Carstanjen, H. D. & Trebin, H.-R. Quasicrystalline d-AlCuCo identified as random tiling by ion channeling combined with particle-induced X-ray emission. *Phys. Rev. Lett.* **82**, 5273–5276 (1999).
16. Pennycook, S. J. & Jesson, D. E. High-resolution Z-contrast imaging of crystals. *Ultramicroscopy* **37**, 14–38 (1991); Atomic-resolution Z-contrast imaging of interfaces. *Acta Metall. Mater.* **40**, S149–S159 (1992).
17. Muller, D. A., Edward, B., Kirkland, E. J. & Silcox, J. Simulation of thermal diffuse scattering including a detailed phonon dispersion curve. *Ultramicroscopy* **86**, 371–380 (2001).
18. Ritsch, S. *et al.* Highly perfect decagonal Al-Co-Ni quasicrystal. *Phil. Mag. Lett.* **74**, 99–106 (1996).
19. Saitoh, K. *et al.* Structural study of an $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ decagonal quasicrystal using the high-angle

- annular dark-field method. *Jpn. J. Appl. Phys.* **36**, L1400–L1402 (1997).
20. Yan, Y., Pennycook, S. J. & Tsai, A. P. Direct imaging of local chemical disorder and columnar vacancies in ideal decagonal Al-Ni-Co quasicrystals. *Phys. Rev. Lett.* **81**, 5145–5148 (1998).
21. Steinhardt, P. J. *et al.* Experimental verification of the quasi-unit-cell model of quasicrystal structure. *Nature* **396**, 55–57 (1998); correction *Nature* (**399**), 84 (1999).
22. Gummelt, P. Construction of Penrose tilings by a single aperiodic protoset. *Geom. Dedicata* **62**, 1–17 (1996).
23. Abe, E. *et al.* Quasi-unit cell model for an Al-Ni-Co ideal quasicrystal based on clusters with broken tenfold symmetry. *Phys. Rev. Lett.* **84**, 4609–4612 (2000).
24. Yan, Y. & Pennycook, S. J. Chemical ordering in $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ decagonal quasicrystals. *Phys. Rev. Lett.* **86**, 1542–1545 (2001).
25. Abe, H. *et al.* Anomalous Debye-Waller factor associated with an order-disorder transformation in an $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ decagonal quasicrystal. *J. Phys.* (submitted).
26. Henley, C. L., Mihalkovic, M. & Widom, M. Total-energy-based prediction for d(AlNiCo). *J. Alloys Comp.* **342**, 221–227 (2002).
27. Takakura, H., Yamamoto, A. & Tsai, A. P. The structure of decagonal $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ quasicrystal. *Acta Crystallogr. A* **57**, 576–585 (2001).
28. Cervellino, A., Haibach, T. & Steurer, W. Structure solution of the basic decagonal Al-Co-Ni phase by the atomic surfaces modeling method. *Acta Crystallogr. B* **58**, 8–33 (2002).
29. Hiraga, K., Ohsuna, T. & Nishimura, S. An ordered arrangement of atom columnar clusters in a pentagonal quasiperiodic lattice of an Al-Ni-Co decagonal quasicrystal. *Phil. Mag. Lett.* **80**, 653–659 (2000).
30. Weickenmeier, A. & Kohl, H. Computation of absorptive form factors for high-energy electron diffraction. *Acta Crystallogr. A* **47**, 590–597 (1991).

Acknowledgements We thank H. Takakura, T. J. Sato, N. Tanaka, K. Ishizuka, M. Widom and C. L. Henley for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.A. (e-mail: abe.eiji@nims.go.jp).

Photocontrolled reversible release of guest molecules from coumarin-modified mesoporous silica

Nawal Kishor Mal, Masahiro Fujiwara & Yuko Tanaka

Kansai Center, National Institute of Advanced Industrial Science and Technology (AIST-Kansai), Ikeda, Osaka 563-8577, Japan

Since the discovery¹ of MCM-41 more than ten years ago, many investigations have explored the suitability of hexagonal mesoporous silicas for potential practical applications^{2–4}. These range from catalysis^{5,6} and optically active materials^{7,8} to polymerization science^{9–12}, separation technology^{3,13,14} and drug delivery^{15–18}, with recent successes in the fabrication of hybrid mesoporous organosilicas^{19–21} expected to open up further application possibilities. Because the pore voids of this class of materials exhibit relatively narrow pore size distributions in the range of 2–4 nm in diameter, mesoporous silicas can selectively include organic compounds and release them continuously at a later stage. The functionalization of MCM-41 pore voids with photoactive derivatives^{22–25} provides influence over the material's absorption behaviour, but full control over the release process remains difficult. Here we show that the uptake, storage and release of organic molecules in MCM-41 can be regulated through the photocontrolled and reversible intermolecular dimerization^{26,27} of coumarin derivatives attached to the pore outlets. Successful functionalization requires uncalcined MCM-41 still filled with the template molecules that directed the formation of its pores, to ensure that coumarin derivatives attach preferentially to the pore outlets, rather than their inside walls. We find that this feature and the one-dimensional, isolated nature of the individual pores allow for efficient and reversible photocontrol over guest access to the material's interior.

The preparation of MCM-41 modified with a substituent bearing a coumarin group used as-synthesized MCM-41 and 7-[(3-triethoxysilyl)propoxy]coumarin (for full details, see Methods). Sample **1A** consists of MCM-41 with 4.0 wt% coumarin substituent. The molecular length of 7-[(3-triethoxysilyl)propoxy]coumarin, attached to silanol groups on MCM-41, is approximately 1.3 nm (see Supplementary Information); dimerization should yield cyclobutane coumarin dimers in anti head-to-head configuration and sufficiently long²⁷ to obstruct the entrances to the pore of MCM-41 (pore diameter ~3 nm).

The ultraviolet (UV) diffuse reflectance spectrum of coumarin-modified MCM-41 (sample **1A**) is shown as a bold line in Fig. 1a. The absorption of the coumarin part (wavelength of maximum absorption $\lambda_{\text{max}} = 324$ nm) is clearly visible, confirming the successful grafting of coumarin substituents. Solid-state ²⁹Si NMR spectra also confirm grafting of organic molecules to the silica (Supplementary Information).

Irradiation of sample **1A** with UV light with wavelengths longer than 310 nm induces photodimerization, and thus results in a gradual decrease in the absorption band centred at 324 nm; the adsorption has almost disappeared after 50 min (Fig. 1a). In contrast, irradiation with UV light with a wavelength around 250 nm regenerates the coumarin absorption band within 2–3 min, owing to photocleavage of the coumarin dimers; however, the absorption feature decreases again under prolonged irradiation (Fig. 1b).

We regard the coumarin substituents as ‘hinged double doors’ controlling access to the pores of MCM-41. In the absence of any photo-irradiation and after photocleavage of the dimers (when the coumarin substituents are present as monomers), the ‘double doors’ are open and allow access to the interior of MCM-41. After photodimerization, the double doors are closed, with cyclobutane dimers spanning the pore entrances and obstructing access—the pore void is isolated. In contrast, the reversible photochromism effects reported for diazo^{8,23} and spiropyran derivatives^{24,25} attached to mesoporous materials are not expected to regulate pore access: in many cases, the optically active molecules are incorporated inside the pores, and the reactions are intramolecular and thus give rise to significantly smaller changes in the extent of pore obstruction.

Controlled-release experiments are conducted with the steroid cholestane, because its molecular size (~1.9 nm length and ~0.6 nm diameter)²⁸ allows it to be stored in the pores of MCM-41. In addition, cholestane has no UV absorption band

around 324 nm that would interfere with monitoring of coumarin spectra. The amount of cholestane absorbed in the pores of sample **2A** (MCM-41 modified with coumarin derivatives and exposed to light after guest entrance, to induce dimerization) was 28 wt% (see Methods and Table 1). Washing in the absence of irradiation resulted in no cholestane remaining in the modified MCM-41. Unmodified MCM-41 also failed to retain cholestane after washing. Similarly, photo-induced dimerization of the coumarin-modified MCM-41 (sample **1A**) before exposure to guest solution resulted in no cholestane being absorbed in the material. These observations suggest that cholestane is retained in modified MCM-41 owing to cyclobutane coumarin dimers preventing passage through the pore outlets.

Sample **2A** was then irradiated with UV light with a wavelength of around 250 nm to cleave the cyclobutane rings, yielding sample **3A**. Washing with *n*-hexane released 77% of the stored cholestane to the solution. To our knowledge, this is the first demonstration of active control over the accessibility of the pores in mesoporous silica.

Table 1 summarizes essential features of the different MCM-41 samples. As indicated, X-ray diffraction measurements (see also Supplementary Information) showed no difference between the MCM-41 samples (samples **1A–3A** and without modification), indicating that the hexagonal silica structure and pore features are not influenced by the coumarin-modification, photo-irradiation, and the cholestane storage/release process. The nitrogen adsorption–desorption isotherms (Fig. 2), however, exhibit distinct differences: whereas coumarin-modification has no influence on nitrogen uptake and release by sample **1A**, coumarin photodimerization (yielding sample **2A**) and subsequent photocleavage (yielding sample **3A**) influence these properties (Fig. 2; for pore size distribution data, see Supplementary Information). As illustrated in the figure, cholestane storage in the pores of the photodimerized sample leads to a very marked decrease in the amount of adsorbed nitrogen (sample **2A**), with the specific surface area and pore volume reduced to a quarter of the values seen for sample **1A**. UV-induced ring cleavage and washing with *n*-hexane, to give sample **3A**, increases the specific surface area as well as the pore volume, as expected if the washing treatment releases most of the cholestane from the MCM-41 pores to the *n*-hexane solution. This process is complete after 48 h (releasing rate; see Supplementary Information). Because the photocleavage of the coumarin dimers proceeds imperfectly^{26,27}, some cholestane is expected (and is

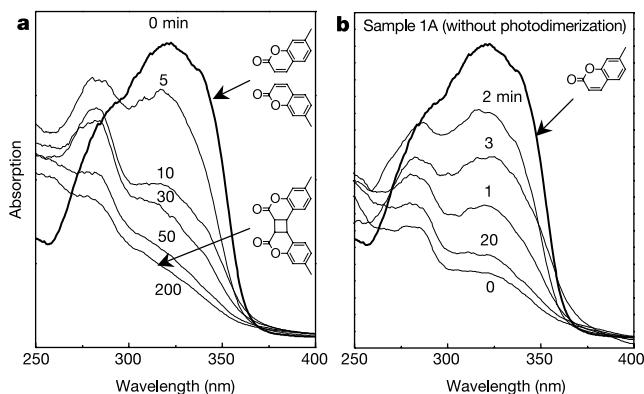


Figure 1 Changes in UV–visible spectra of modified MCM-41 samples during UV irradiation. **a**, Sample **1A** (MCM-41 with 4 wt% coumarin substituent, prepared with a grafting time of 15 min; see Methods) during photo-irradiation with UV light of wavelength >310 nm. **b**, Sample **2A** (cholestane absorbed to sample **1A**, followed by photodimerization and washing) during photo-irradiation with UV light of wavelength around 250 nm. The structures of coumarin monomer and dimer are also shown, and the spectra corresponding to the presence of only monomers or dimers are identified.

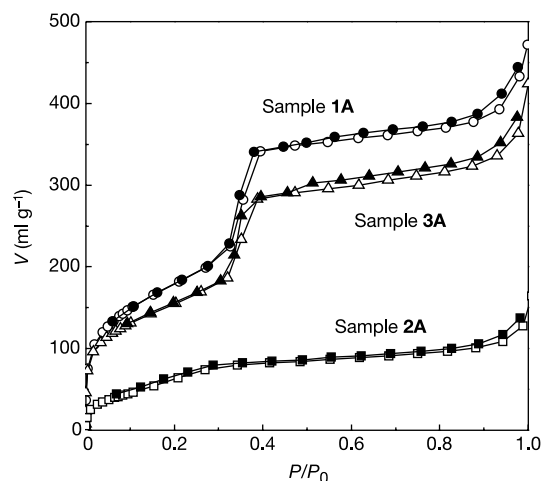


Figure 2 Nitrogen adsorption–desorption isotherms of MCM-41 samples. Shown are data for sample **1A** (see Fig. 1 legend), sample **2A** (see Fig. 1 legend) and sample **3A** (photocleaved and washed sample **2A**). *V*, volume of N₂ adsorbed; *P*, pressure of adsorbent; *P*₀, saturation pressure (1 atm). Open symbols, adsorption branch of isotherm; closed symbols, desorption branch of isotherm.

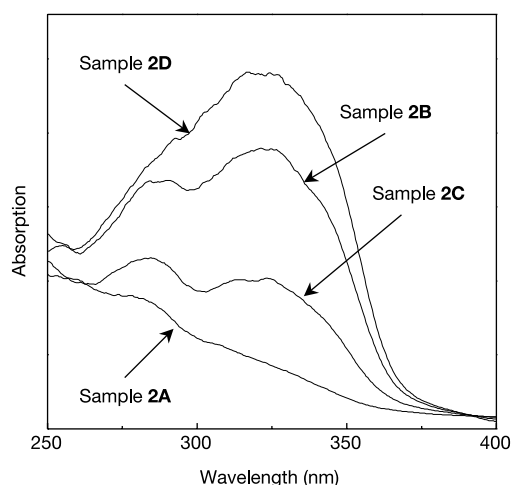


Figure 3 UV-visible spectra of coumarin-modified MCM-41 after complete photo-irradiation by UV light of wavelength >310 nm. Data are shown for sample **2A** (see Fig. 1 legend), sample **2B** (cholestane absorbed, photodimerized and washed 5.6 wt% coumarin grafted MCM-41; grafting time, 22 h), sample **2C** (cholestane absorbed,

photodimerized and washed 4 wt% coumarin grafted on calcined MCM-41 for 15 min) and sample **2D** (cholestane absorbed, photodimerized and washed 5.6 wt% coumarin grafted MCM-41 produced by one-pot synthesis).

observed) to remain in the pores (Table 1). Comparable uptake and release behaviour was observed with pyrene and phenanthrene, molecules that are smaller than cholestane. Similarly, the uptake and release of progesterone, a corpus luteum hormone, could also be controlled (after photodimerization: 28.2 wt% storage, and after photocleavage: 6.0 wt% storage).

To achieve effective release control, the coumarin substituents need to be attached to the pore outlets of the modified MCM-41. This was achieved by modifying as-synthesized MCM-41 at ambient temperature (~ 295 K) over a short grafting time (15 min). When using identical conditions except for a grafting time of 22 h (sample **1B**; 5.6 wt% grafted), the material exhibited significantly worse performance in controlled uptake and release experiments (Table 1, samples **2B** and **3B**). Similarly, when an equal amount of coumarin substituent was grafted in a short period (15 min) to calcined MCM-41 (sample **1C**; 4.0 wt% grafted), cholestane uptake was significantly diminished (Table 1, samples **2C** and **3C**). Finally, coumarin-modified MCM-41 was also prepared by one-pot hydrothermal synthesis (sample **1D**; 5.6 wt% grafted), but showed no evidence for controlled release after extraction of the surfactant, or of cholestane storage after photo-irradiation (Table 1, sample **2D**).

Figure 3 gives the UV spectra of samples **2A**, **2B**, **2C** and **2D** after

photo-irradiation (at wavelengths >310 nm) over 200 min. In the case of sample **2A**, the coumarin absorption at 324 nm has almost disappeared, indicating complete dimerization of coumarin ($>98\%$). In the case of sample **2B** and **2C**, the less-pronounced decrease in the absorption at 324 nm indicates poor (25%) and incomplete (63%) dimerization, respectively, while sample **2D** shows essentially no evidence for photodimerization ($<1\%$). These results indicate that, as expected, the position of the coumarin substituents in the pore of MCM-41 determines their reactivity to photodimerization. That is, the crowding of coumarin substituents at the pore outlets in sample **1A** enables almost complete photodimerization, whereas the more dispersed attachment in samples **2B** and **2C** allows for only limited coumarin photodimerization. In the case of sample **2B**, which used as-synthesized MCM-41, the long grafting time permits gradual displacement of surfactants and diffusion of coumarin reagents into the inner pore space. (More than 90% of the template was removed after this procedure). In the case of sample **2C**, grafting²⁹ on calcined MCM-41 but only over a short period resulted in preferential attachment near the outlet of the pores^{30,31}. We expect that the large size of the coumarin substituents that we used, and their associated slow diffusion, will have enhanced this tendency, thus explaining the overall better

Table 1 Properties of various modified MCM-41 samples

Sample	Status	d_{100}^* (nm)	$S_{\text{BET}}^\dagger$ ($\text{m}^2 \text{g}^{-1}$)	$V_p^\ddagger$ ($\text{cm}^3 \text{g}^{-1}$)	APD § (nm)	Peak pore diameter $^\parallel$ (nm)	Cholestane storage (wt%)
MCM-41	Unmodified	4.06	982	0.84	2.80	3.0	0.0
Sample 1A #	Modified	4.06	970	0.82	2.79	3.0	0.0
Sample 2A	Dimerized	4.06	275	0.21	1.99	1.9	28.0
Sample 3A	Cleaved	4.06	747	0.61	2.60	2.8	6.4
Sample 1B ☆	Modified	4.06	907	0.74	2.71	2.7	0.0
Sample 2B	Dimerized	4.06	790	0.58	2.43	2.6	4.8
Sample 3B	Cleaved	4.06	842	0.70	2.69	2.8	1.4
Sample 1C **	Modified	4.04	916	0.77	2.74	2.8	0.0
Sample 2C	Dimerized	4.04	615	0.49	2.42	2.3	12.8
Sample 3C	Cleaved	4.04	798	0.70	2.72	2.8	2.1
Sample 1D ‡‡	Modified	3.95	853	0.62	2.38	2.2	0.0
Sample 2D	Dimerized	3.95	856	0.62	2.38	2.2	<0.1

* d_{100} : X-ray diffraction (100) interplanar spacing.

† S_{BET} : BET specific surface area.

‡ V_p : Primary mesopore volume.

§APD: average pore diameter = $4V_p/S_{\text{BJH}}$, where S_{BJH} is the BJH specific surface area.

||Peak pore diameter calculated from BJH pore size distribution curve using adsorption branches.

¶Weight per cent of cholestane stored in the modified MCM-41 materials after thorough washing with *n*-hexane.

#Sample **A**: 4 wt% coumarin grafted MCM-41; grafting time, 15 min.

☆Sample **B**: 5.6 wt% coumarin grafted MCM-41; grafting time, 22 h.

Sample **C: 4 wt% coumarin grafted on calcined MCM-41 for 15 min.

‡‡Sample **D**: 5.6 wt% coumarin grafted in one-pot synthesis.

controlled-release performance of sample **2C**, relative to that of sample **2B** (Table 1). Of course, we expect that the entire external surface of all these modified MCM-41 samples has been functionalized during treatment with the coumarin derivatives, but this should not affect the performance of the materials in our controlled-release experiments. Finally, in the modified MCM-41 sample produced by one-pot synthesis (sample **2D**), the coumarin substituents are expected to exist randomly in the pore space, with this highly dispersed placement of the substituents preventing photodimerization of coumarin groups and the storage of chemicals in the pore.

The use of as-synthesized MCM-41 and a short time of grafting are essential for preparing materials that allow effective and direct control over guest uptake and release. Porous materials with interconnected pore architectures, such as silica gels, are expected to be less effective, because guest molecules could diffuse to the outside through remaining unmodified (not closed) pore outlets. Indeed, controlled release using coumarin-modified silica gel was not successful: although the photodimerization took place, no cholestane was stored in the gel. The present work shows that the isolated, one-dimensional cylindrical pores of MCM-41 render this material particularly suitable for controlled-release applications. □

Methods

Preparation of the coumarin-modified MCM-41

MCM-41 was prepared using gel of the following molar composition; 1 SiO₂: 0.51 CTMABr: 0.67 TMAOH: 0.46 Na₂O: 0.65 H₂SO₄: 80 H₂O (pH 9.75). The gel was heated at 373 K for 3 d. The obtained material was filtered off, washed and dried at 373 K for 24 h. 2 g of as-synthesized MCM-41 was suspended in a solution containing 20 ml of *n*-hexane and 0.20 g of 7-[(3-triethoxysilyl)propoxy]coumarin under stirring at ambient temperature for 15 min. *n*-Hexane was evaporated by a rotary evaporator at 353 K for 2 h, and the resulting material was dried under vacuum at 423 K for 12 h. 2 g of this coumarin-modified as-synthesized MCM-41 with surfactant was refluxed in 100 ml of ethanol containing 4 ml of HCl (1 M) at 353 K for 4 h. The solid was descended and washed with ethanol. This process was carried out twice to ensure the complete removal of surfactant from the pores of MCM-41. The obtained solid was filtered off, washed with ethanol and water, and finally dried at 353 K for 12 h. This material was referred as sample **1A**. Sample **1B** was prepared by the same grafting procedure as sample **1A** for 22 h duration. Sample **1C** was obtained by grafting 4.0 wt% of coumarin substituent on calcined MCM-41 at ambient temperature in hexane solution under stirring for 15 min. Sample **1D** was synthesized from the mixed solution of tetraethoxysilane and 7-[(3-triethoxysilyl)propoxy]coumarin (5 mol% to tetraethoxysilane) by the same procedure as mentioned above. The removal of surfactant and the amount of substituents in MCM-41 were assessed by gas chromatography, thermogravimetric analysis, and chemical analysis.

Preparation of 7-[(3-triethoxysilyl)propoxy]coumarin

7-Allyloxycoumarin was obtained from umbelliferone and allyl bromide in the presence of potassium carbonate, and purified by recrystallization from ethanol. The yield was over 90%. 7-[(3-triethoxysilyl)propoxy]coumarin was prepared as follows; after bubbling dry nitrogen into a toluene solution (100 ml) of 7-allyloxycoumarin 3.24 g (16 mmol) and triethoxysilane 2.94 g (17.6 mmol) for 10 min, 0.8 ml of a toluene solution (2 mM) of Pt [(dvs) [platinum(0)-1,3-divinyl-1,1,3,3-tetramethyldisiloxane complex] (Aldrich) was added, and the resulting solution was stirred for 12 h at room temperature. After removing solvent under reduced pressure, the oil obtained (successfully identified as 7-[(3-triethoxysilyl)propoxy]coumarin) was used directly to modify MCM-41.

Photoresponsive on-off controlled-release experiment

1 g of coumarin-modified MCM-41 (samples **1A**, **1B**, **1C** and **1D**) was suspended in an *n*-hexane solution (20 ml) containing 1 g of cholestane at ambient temperature for 24 h. The resulting solid was filtered off, washed with *n*-hexane many times on a filter, and dried at 333 K for 12 h. The content of cholestane in the filtrate was analysed using gas chromatography. The obtained solid was photodimerized by irradiation with UV light of wavelength >310 nm for 30 min using a 450-W high-pressure mercury lamp through a Pyrex glass cooler. After the dimerization, the solid was suspended in *n*-hexane (100 ml) and stirred at room temperature for 48 h. Finally, the solid was filtered off, washed with *n*-hexane and dried at room temperature (samples **2A**, **2B**, **2C** and **2D**). The amounts of cholestane in samples **2A**, **2B**, **2C** and **2D** were estimated from the difference between the initial and the recovered amounts. This dimerized material was subjected to photocleavage by irradiation at around 250 nm wavelength for 2.5 min with a low-pressure mercury lamp through a quartz glass cooler, and treated with *n*-hexane (100 ml) at room temperature under stirring for 48 h. Finally, the solid was descended, washed and dried at room temperature (samples **3A**, **3B** and **3C**). The amounts of cholestane released

(dissolved in *n*-hexane) were analysed using gas chromatography and thermogravimetric analysis.

Received 7 June; accepted 10 December 2002; doi:10.1038/nature01362.

- Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. Ordered mesoporous molecular sieves synthesized by a liquid-crystal template mechanism. *Nature* **359**, 710–712 (1992).
- Ying, J. Y., Mehnert, C. P. & Wong, M. S. Synthesis and applications of supramolecular-templated mesoporous materials. *Angew. Chem. Int. Edn Engl.* **38**, 56–77 (1999).
- Stein, A., Melde, B. J. & Schrodin, R. C. Hybrid inorganic-organic mesoporous silicates—Nanoscale reactors coming of age. *Adv. Mater.* **12**, 1403–1419 (2000).
- Davis, M. E. Ordered porous materials for emerging applications. *Nature* **417**, 813–821 (2002).
- Maschmeyer, T., Rey, F., Sanker, G. & Thomas, J. M. Heterogeneous catalysts obtained by grafting metallocene complexes onto mesoporous silica. *Nature* **378**, 159–162 (1995).
- Corma, A. From microporous to mesoporous molecular sieve materials and their use in catalysis. *Chem. Rev.* **97**, 2373–2419 (1997).
- Nguyen, T. Q., Wu, J. J., Doan, V., Schwartz, B. J. & Tolbert, S. H. Control of energy transfer in oriented conjugated polymer-mesoporous silica composites. *Science* **288**, 652–656 (2000).
- Ganschow, M., Wark, M., Wohrle, D. & Schulz-Ekloff, G. Anchoring of functional dye molecules in MCM-41 by microwave-assisted hydrothermal co-condensation. *Angew. Chem. Int. Edn Engl.* **39**, 161–163 (2000).
- Wu, C. G. & Bein, T. Conducting polyaniline filaments in a mesoporous channel host. *Science* **264**, 1757–1759 (1994).
- Zhou, W. Z. *et al.* Ordering of ruthenium cluster carbonyls in mesoporous silica. *Science* **280**, 705–708 (1998).
- Kageyama, K., Tamazawa, J. & Aida, T. Extrusion polymerization: Catalyzed synthesis of crystalline linear polyethylene nanofibers within a mesoporous silica. *Science* **285**, 2113–2115 (1999).
- Spange, S. *et al.* Cationic host-guest polymerization of *N*-vinylcarbazole and vinyl ethers in MCM-41, MCM-48, and nanoporous glasses. *Chem. Eur. J.* **7**, 3722–3728 (2001).
- Feng, X. *et al.* Functionalized monolayers on ordered mesoporous supports. *Science* **276**, 923–926 (1997).
- Mercier, L. & Pinnavaia, T. J. Access in mesoporous materials: Advantages of a uniform pore structure in the design of a heavy metal ion adsorbent for environmental remediation. *Adv. Mater.* **9**, 500–503 (1997).
- Ahola, M., Korteso, P., Kangasniemi, I., Kiesvaara, J. & Antti Yli-Urpo, A. Silica xerogel carrier material for controlled release of toremifene citrate. *Int. J. Pharm.* **195**, 219–227 (2000).
- Arcos, D., Ragel, C. V. & Vallet-Regí, M. Bioactivity in glass/PMMA composites used as drug delivery system. *Biomaterials* **22**, 701–708 (2001).
- Czuryszkiewicz, T. *et al.* Drug release from biodegradable silica fibers. *J. Non-Cryst. Solids* **306**, 1–10 (2002).
- Vallet-Regí, M., Rámila, A., del Real, R. P. & Pérez-Pariente, J. A new property of MCM-41: Drug delivery system. *Chem. Mater.* **13**, 308–311 (2001).
- Asefa, T., MacLachlan, M. J., Coombs, N. & Ozin, G. A. Periodic mesoporous organosilicas with organic groups inside the channel walls. *Nature* **402**, 867–871 (1999).
- Fujiwara, M. *et al.* Preparation of MCM-41/naion composite material; a selective catalyst of α -methylstyrene dimerization. *J. Chem. Soc. Chem. Commun.* 1523–1524 (2000).
- Inagaki, S., Guan, S., Ohsuna, T. & Terasaki, O. An ordered mesoporous organosilica hybrid material with a crystal-like wall structure. *Nature* **416**, 304–307 (2002).
- Alvaro, M., Ferrer, B., García, H. & Rey, F. Photochemical modification of the surface area and tortuosity of a *trans*-1,2-bis(4-pyridyl)ethylene periodic mesoporous MCM organosilica. *J. Chem. Soc. Chem. Commun.* 2012–2013 (2002).
- Hoffmann, K., Resch-Genger, U. & Marlow, F. Photoinduced switching of nanocomposites consisting of azobenzene and molecular sieves: Investigation of the switching states. *Micropor. Mesopor. Mater.* **41**, 99–106 (2000).
- Wirmsberger, G., Scott, B. J., Chmelka, B. F. & Stucky, G. D. Fast response photochromic mesostructures. *Adv. Mater.* **12**, 1450–1454 (2000).
- Schomburg, C., Wark, M., Röhling, Y., Schulz-Ekloff, G. & Wöhrle, D. Photochromism of spiropyran in molecular sieve voids: Effects of host-guest interaction on isomer status, switching stability and reversibility. *J. Mater. Chem.* **11**, 2014–2021 (2001).
- Chujo, Y., Sada, K. & Saegusa, T. Polyoxyazoline having a coumarin moiety as a pendant group. Synthesis and photogelation. *Macromolecules* **23**, 2693–2697 (1990).
- Li, W. J., Lynch, V., Thompson, H. & Fox, M. A. Self-assembled monolayers of 7-(10-thiododecoxy)coumarin on gold: Synthesis, characterization, and photodimerization. *J. Am. Chem. Soc.* **119**, 7211–7217 (1997).
- Berezin, M. Y., Dzenitis, J. M., Hughes, B. M. & Ho, S. V. Separation of sterols using zeolites. *Phys. Chem. Chem. Phys.* **3**, 2184–2189 (2001).
- Kruk, M., Antochshuk, V., Matos, J. R., Mercuri, L. P. & Jaroniec, M. Determination and tailoring the pore entrance size in ordered silicas with cage-like mesoporous structures. *J. Am. Chem. Soc.* **124**, 768–769 (2002).
- Zhao, X. S., Lu, G. Q. M. & Hu, X. A novel method for tailoring the pore-opening size of MCM-41 materials. *J. Chem. Soc. Chem. Commun.* 1391–1392 (1999).
- Lim, M. H. & Stein, A. Comparative studies of grafting and direct syntheses of inorganic-organic hybrid mesoporous materials. *Chem. Mater.* **11**, 3285–3295 (1999).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements N.K.M. was supported by an STA fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.F. (e-mail: m-fujiwara@aist.go.jp).

Abrupt changes in the Asian southwest monsoon during the Holocene and their links to the North Atlantic Ocean

Anil K. Gupta*, David M. Anderson† & Jonathan T. Overpeck‡

* Department of Geology & Geophysics, Indian Institute of Technology, Kharagpur 721 302, India

† NOAA Paleoclimatology Program, 325 Broadway, Code E/GC, Boulder, Colorado 80305-3328, USA

‡ Center for the Study of Planet Earth and the Department of Geosciences, University of Arizona, Tucson, Arizona 85721, USA

During the last ice age, the Indian Ocean southwest monsoon exhibited abrupt changes that were closely correlated with millennial-scale climate events in the North Atlantic region^{1–3}, suggesting a mechanistic link. In the Holocene epoch, which had a more stable climate, the amplitude of abrupt changes in North Atlantic climate was much smaller, and it has been unclear whether these changes are related to monsoon variability. Here we present a continuous record of centennial-scale monsoon variability throughout the Holocene from rapidly accumulating and minimally bioturbated sediments in the anoxic Arabian Sea. Our monsoon proxy record reveals several intervals of weak summer monsoon that coincide with cold periods documented in the North Atlantic region⁴—including the most recent climate changes from the Medieval Warm Period to the Little Ice Age and then to the present. We therefore suggest that the link between North Atlantic climate and the Asian monsoon is a persistent aspect of global climate.

Abrupt changes in climate are apparent in the geologic record of climate over various timescales. Examples of large millennial-scale abrupt events in the North Atlantic include the Dansgaard–Oeschger cycles of the last glacial, and the Younger Dryas event that occurred at the end of that glacial. Because these North Atlantic warm/cold cycles correlate with monsoon variations^{2,3,5}, a question relevant to future climate change and prediction efforts is whether the smaller abrupt changes that have occurred in the North Atlantic during the present Holocene interglacial period are also reflected in the monsoon. The Indian Ocean monsoon system is an important

component of the global climate, and has a significant role in the socio-economic life of people of the Indian subcontinent. Seasonal reversals in the Indian monsoons, the southwest (or ‘summer’) and northeast (or ‘winter’) monsoon, influence weather and climate between 30° N and 30° S over African, Indian and Asian landmasses. Although evidence of the monsoon can be found across the region in various proxies, arguably the most consistent continuous record of the monsoon comes from the floor of the Arabian Sea.

Each summer adjacent to the Oman margin, the seasonal reversal in the winds causes southwest winds to blow across the Arabian Sea (Fig. 1). Cooling of the sea surface associated with coastal and open-ocean upwelling promotes the blooming of distinct fauna and flora^{6–8}. The biological response to the monsoonal activity in the surface water column is preserved as increased abundance of the planktic foraminifer *Globigerina bulloides*. The *G. bulloides* proxy has been calibrated using modern sea-floor samples⁶ and sediment-trap time series⁹, and has been tested over a range of timescales. Advantages of this proxy are (1) its unique association with the summer monsoon (*G. bulloides* has a subpolar habitat and would be absent in the tropics except for wind-driven upwelling), (2) linear correlation with the surface cooling due to upwelling, apparently unbiased by other influences, and (3) strong sensitivity to wind speed and the monsoonal atmospheric pressure gradient (the wind

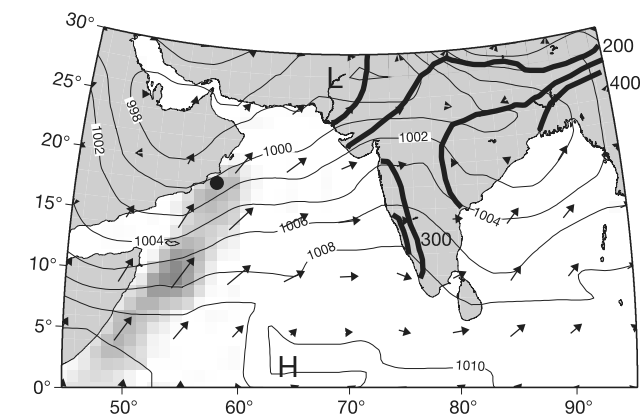


Figure 1 July sea-level pressure (mbar, thin contours), wind direction²⁷, cooling of the Arabian Sea due to upwelling²⁸, and precipitation over Asia²⁹ (mm month^{−1}, thick contours). Site 723 and box core RC2730 are located at 18° N, 58° E (circle). H, high pressure; L, low pressure.

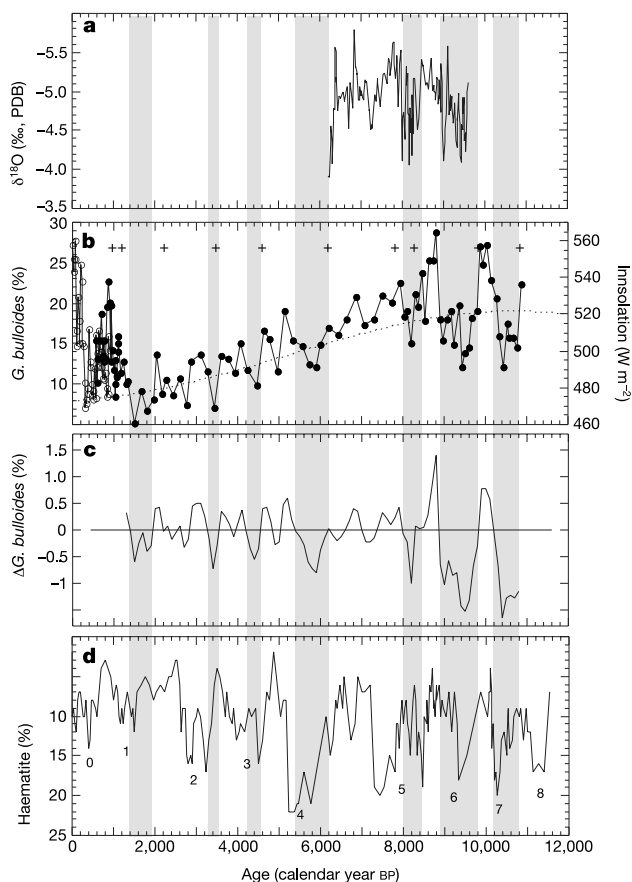


Figure 2 Southwest monsoon proxy record from the Arabian Sea Site 723A and box core RC2730 combined with Oman cave stalagmite $\delta^{18}\text{O}$ and North Atlantic haematite percentage. Time series of **a**, cave stalagmite $\delta^{18}\text{O}$ from ref. 16, **b**, *G. bulloides* percentage in Hole 723A (filled circles) and box core RC2730 (open circles) from the Arabian Sea (detail in Fig. 3), and July insolation at 65° N (ref. 18, shown by dotted line; radiocarbon-dated intervals shown by crosses), **c**, change in *G. bulloides* percentage (normalized by removing the trend related to insolation), and **d**, haematite percentage in core MC52–VM29–191 from the North Atlantic, and events labelled 0–8 in ref. 4. The vertical grey bars indicate intervals of weak Asian SW monsoon.

Table 1 Radiocarbon ages determined by accelerator mass spectrometry

NOSAMS* lab. Code	Species	Depth (m.b.s.f.)†	Radiocarbon age‡ (yr BP)	Calendar age (years before 1950)	1 s.d. (yr)
34259	Mixed planktics	0.16	1,680	1,005	35
34260	<i>G. bulloides</i> & <i>N. dutertrei</i>	0.39	2,010	1,330	45
34261	<i>G. bulloides</i> & <i>N. dutertrei</i>	0.64	2,760	2,254	30
34262	<i>G. bulloides</i> & <i>N. dutertrei</i>	0.96	3,890	3,587	45
34263	<i>G. bulloides</i>	1.21	4,670	4,622	40
34264	<i>G. bulloides</i>	1.54	6,000	6,202	30
34265	<i>G. bulloides</i>	1.54	6,070	6,276	30
34266	<i>G. bulloides</i>	1.86	7,700	7,938	45
34267	<i>G. bulloides</i>	2.18	8,110	8,359	40
34268	<i>G. bulloides</i>	2.74	9,510	9,854	40
34269	<i>G. bulloides</i>	3.16	10,500	10,877	40

*NOSAMS, National Ocean Sciences Accelerator Mass Spectrometer Facility.

†Metres below sea floor.

‡Calibrated to calendar year before present (1950) using the marine98 calibration data²⁰.

stress that drives upwelling is proportional to the square of the wind speed).

Like other major components of climate, the summer monsoon wind has varied at orbital^{7,10–12} and longer timescales in response to known external forcing. More remarkable are the abrupt changes that have occurred in the absence of known forcing over suborbital or millennial scales^{1,5,8}. Millennial-scale abrupt monsoon events have been attributed to changes in the North Atlantic and downstream over Europe and Eurasia, according to the hypothesis that increased winter snowfall weakens the monsoon the following summer, observed in numerical simulations and observation-based studies^{1,13,14}. The recent discovery that the Holocene North Atlantic is also marked by repeating, but less severe, cold spells^{4,15} (the most recent being the Medieval Warm Period and subsequent Little Ice Age) begs the question: were the small-amplitude changes

in the North Atlantic during the Holocene also accompanied by changes in the Asian southwest monsoon? Pioneering studies have suggested such a link for the Asian southwest monsoon¹⁶ and the East Asian monsoon¹⁷, but more records are needed, given the uncertainty in radiocarbon dating (± 100 yr), the effect of bioturbation on marine sediments, the generally noisy character of most proxies, and the small amplitude of the expected signal (barely above the uncertainty). To determine whether the Little Ice Age, Medieval Warm Period and other Holocene events were accompanied by changes in the Asian southwest monsoon, we examined one of the most extensively tested proxies of the Asian southwest monsoon, the record of *G. bulloides* shells in sediments from the Oman margin.

We produced a record of *G. bulloides* covering 10,877 calendar years by sampling cores from Ocean Drilling Program (ODP) Site 723, Hole A, every 1 cm (0–0.25 m), and every 4 cm (0.25–3.16 m) (see Supplementary Information). Site 723A is located on the continental margin of Oman, Arabian Sea ($18^{\circ}03.079'N$, $57^{\circ}36.561'E$; water depth 807.8 m) in the centre of an oxygen-minimum zone, and provides a high-resolution sedimentary record of biotic variations linked with the intense southwest monsoon upwelling (Fig. 1). The bioturbation smoothing that would normally affect fine-scale variability is minimal at this site owing to the strong oxygen-minimum zone that exists in the Oman margin. The average age interval per sample is ~ 133 yr (ranging from 53 to 166 yr) based on linear interpolation of ten AMS ^{14}C calibrated dates (Table 1; also see Supplementary Information). Although the radiocarbon measurement error (1 s.d.) is less than 50 yr, including the uncertainty in reservoir age and stratigraphic interpolation increases the uncertainty to about ± 100 yr. To determine the age of the uppermost sample, we correlated the top of our record with a well-dated Oman margin box core (RC2730), finding an age of ~ 560 yr, suggesting that the top few centimetres were lost at the time of drilling. The percentages of *G. bulloides* were calculated from an aliquot of ~ 300 specimens of planktic foraminifera from the $>150\text{-}\mu\text{m}$ size fraction from each sample. The *G. bulloides* percentage data are compared with the $65^{\circ}N$ July solar insolation¹⁸ and with data on the percentage of haematite-stained grains from the North Atlantic⁴ (Fig. 2).

We identified seven discrete intervals of weak summer monsoon (grey bars, Fig. 2b) during the Holocene that, although small in amplitude, can be correlated within the radiocarbon age uncertainties to the millennial-scale events (cold spells numbered 0–8, Fig. 2d) in the North Atlantic^{4,15} (plotted so that cooling and monsoon weakening are downward in Fig. 2). Monsoon maxima have occurred in several pulses while the North Atlantic was warmest (low haematite). We defined monsoon events as the departures from the Holocene mean (Fig. 2c), removing the variation associated with the slow Holocene change in summer insolation by subtracting the standardized insolation series from the standardized

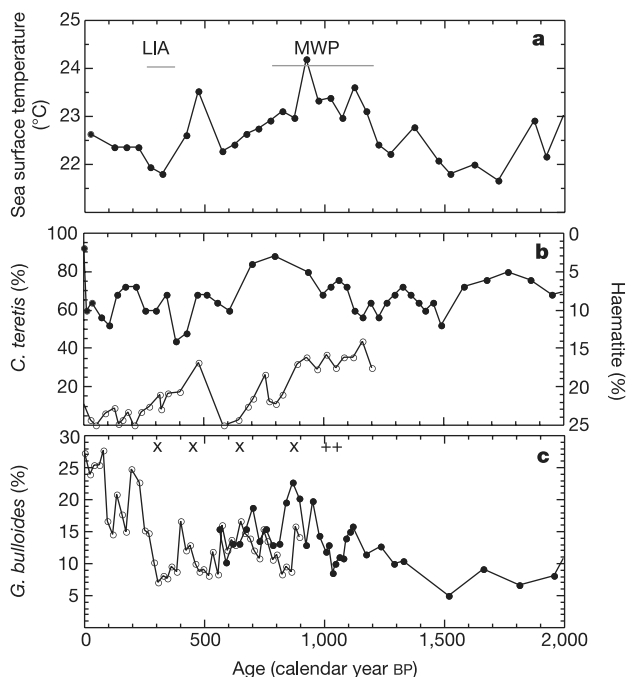


Figure 3 Southwest monsoon proxy record from the Arabian Sea Site 723A and box core RC2730 combined with North Atlantic percentage *Cassidulina teretis*, haematite and SST. Time series of **a**, temperature (inferred from $\delta^{18}O$ variation)²³ and ice rafting events⁴ in North Atlantic sediments, **b**, *C. teretis* abundance from an east Greenland fjord²⁴, and **c**, *G. bulloides* percentage in Hole 723A (filled circles) and box core RC2730 (open circles), showing also dated depths in 723A (plus signs) and RC2730 (crosses). The North Atlantic Medieval Warm Period (MWP) and Little Ice Age (LIA) extents are based on the records shown here.

G. bulloides series. Weak monsoon events were defined as runs in the data exceeding 0.5 s.d. The four oldest events have the largest amplitude, and have readily identifiable counterparts (numbered 7, 6, 5 and 4) in the haematite record. Correlating the weak monsoon events to a record of oxygen isotope enrichment in an Oman cave stalagmite¹⁶ (Fig. 2a), we observe enriched $\delta^{18}\text{O}$ (weaker monsoon/intertropical convergence zone precipitation) accompanying times of weaker winds, as expected. The cool event at 8,200 yr before present (BP), now well-documented in the North Atlantic¹⁹, appears prominent in the cave stalagmite record as well as our monsoon record.

The identification of events in the younger part (5,000 yr BP to present) is more difficult to establish because the events are of short duration and barely exceed 0.5 s.d. Nevertheless, we made tentative correlations with North Atlantic Bond events 3, 2 and 1 (as well as event 0, see Fig. 2b). Potential problems with our data include double-peak event 5 (less prominent in our data), the low amplitude of variability during 2–6,000 yr BP, and the observation that almost all correlations could be improved by small age adjustments. It is important to point out that in addition to the radiocarbon measurement uncertainty (Table 1), the largest source of age error is probably related to small unknown variations in the surface-ocean reservoir age in both the Atlantic Ocean and the Arabian Sea during the Holocene. Despite these problems, the observations provide strong support for the view that throughout the Holocene, cool episodes in the North Atlantic are accompanied by weakening of the Asian southwest monsoon wind, just as they are for the larger, more readily observable Dansgaard–Oeschger cycles of the last glacial.

Variations in solar output have been suggested as the driver of both the North Atlantic cycles⁴ and the variations in Oman precipitation¹⁶. Solar variations could influence the monsoon regime indirectly, perhaps amplified by the North Atlantic teleconnection, on the basis of numerical simulations of North Atlantic cooling^{1,20}. Alternately, solar variations could directly affect the land–sea contrast that drives the monsoon, on the basis of numerical simulations with reduced solar output during the Maunder minimum²¹. These papers^{1,20,21} indicate that the monsoon could be sensitive to relatively small changes in forcing (0.25% change in solar output, or a 2°C change in sea surface temperature). Internal forcing caused by oscillations in North Atlantic northward heat transport/deep-water production are an alternative explanation that excludes the role of insolation. To discriminate among these hypotheses more records are needed, particularly those that quantify the magnitude of sea surface temperature change in the North Atlantic and the magnitude of change in the monsoon.

The North Atlantic Medieval Warm Period and Little Ice Age can be considered to be the most recent in a seemingly persistent series of small-amplitude changes in the Holocene climate of the North Atlantic region. Splicing previously studied box core RC 2730²² with the Site 723 record allows us to examine the last 2,000 yr in greater detail (Fig. 3). We find a broad increase in *G. bulloides* at 600–1,200 yr BP that correlates with a minimum in haematite and with warmer sea surface temperature at the Bermuda rise, North Atlantic²³ (labelled MWP—‘Medieval Warm Period’—in Fig. 3a), and a minimum in *G. bulloides* at approximately 300–400 yr BP that correlates with a brief maximum in haematite and with cooling at Bermuda rise during the Little Ice Age (LIA in Fig. 3a). The monsoon record also correlates with the Medieval Warm Period along the east Greenland margin indicated by increased abundance of *Cassidulina teretis*²⁴ (Fig. 3b). Although not exact, the correlations are unambiguous and allow us to conclude that as for the entire Holocene, the most recent cycles in the North Atlantic have counterparts in the Asian southwest monsoon (Fig. 3c).

The significance of our results lies in demonstrating a pattern of persistent variability in the monsoon throughout the Holocene that may be linked with episodic warming/cooling of the North Atlantic.

These results indicate that subtle interglacial changes in the North Atlantic, and not just large glacial changes, may have a significant effect on the strength of the Asian monsoon system. Our results highlight the need to improve our understanding of abrupt (and difficult to predict) weakening in monsoon strength that could accompany abrupt climate shifts in the North Atlantic that may occur in the future^{25,26}, further complicating efforts to plan for the future in the world’s most populous region. □

Received 13 September; accepted 25 November 2002; doi:10.1038/nature01340.

- Overpeck, J. T., Anderson, D. M., Trumbore, S. & Prell, W. L. The Southwest Monsoon over the last 18,000 years. *Clim. Dyn.* **12**, 213–225 (1996).
- Schulz, H., Rad, U. & Erlenkeuser, H. Correlations between Arabian Sea and Greenland climate oscillations of the past 110,000 years. *Nature* **393**, 54–57 (1998).
- Altabet, M. A., Higginson, M. J. & Murray, D. W. The effect of millennial-scale changes in Arabian Sea denitrification on atmospheric CO₂. *Nature* **415**, 159–162 (2002).
- Bond, G. *et al.* Persistent solar influence on North Atlantic climate during the Holocene. *Science* **294**, 2130–2136 (2001).
- Sirocko, F., Leuschner, D., Staubwasser, M., Maley, J. & Heusser, L. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P., Webb, R. S. & Keigwin, L. D.) 113–126 (American Geophysical Union, Washington DC, 1999).
- Prell, W. L. in *Climate Processes and Climate Sensitivity* (eds Hansen, J. E. & Takahashi, T.) 48–57 (American Geophysical Union, Washington DC, 1984).
- Anderson, D. M. & Prell, W. L. A 300 kyr record of upwelling off Oman during the late Quaternary: Evidence of the Asian Southwest Monsoon. *Paleoceanography* **8**, 193–208 (1993).
- Naidu, P. D. & Malmgren, B. J. A high resolution record of late Quaternary upwelling along the Oman Margin, Arabian Sea based on planktonic foraminifera. *Paleoceanography* **11**, 129–140 (1996).
- Curry, W. B., Ostermann, D. R., Guptha, M. V. S. & Ittekkot, V. in *Upwelling Systems: Evolution since the Early Miocene* (eds Summerhayes, C. P., Prell, W. L. & Emeis, K. C.) 93–106 (The Geological Society, London, 1992).
- Prell, W. L. & Campo, E. V. Coherent response of Arabian Sea upwelling and pollen transport to late Quaternary monsoonal winds. *Nature* **323**, 526–528 (1986).
- Clemens, S. C., Prell, W. L., Murray, D. W., Shimmield, G. & Weedon, G. Forcing mechanisms of the Indian Ocean monsoon. *Nature* **353**, 720–725 (1991).
- Gupta, A. K., Dhirga, H., Melice, J.-C. & Anderson, D. M. Earth’s eccentricity cycles and Indian Summer Monsoon variability over the past 2 million years: evidence from deep sea benthic foraminifer. *Geophys. Res. Lett.* **28**, 4131–4134 (2001).
- Barnett, T. P., Dumenil, L., Schlese, U., Roeckner, E. & Latif, M. The effect of Eurasian snow cover on regional and global climate variations. *J. Atmos. Sci.* **46**, 661–684 (1989).
- Meehl, G. A. Influence of the land surface on the Asian summer monsoon, external conditions versus internal feedbacks. *J. Clim.* **7**, 1033–1049 (1994).
- Bond, G. *et al.* A pervasive millennial-scale cycle in North Atlantic Holocene and Glacial climates. *Science* **278**, 1257–1266 (1997).
- Neff, U. *et al.* Strong coherence between solar variability and the monsoon in Oman between 9 and 6 kyr ago. *Nature* **411**, 290–293 (2001).
- Wang, L. *et al.* Holocene variations in Asian monsoon moisture: a biennial sediment record from the South China Sea. *Geophys. Res. Lett.* **26**, 2889–2892 (1999).
- Berger, A. L. Long-term variations of daily insolation and Quaternary climatic changes. *J. Atmos. Sci.* **35**, 2362–2367 (1978).
- Alley, R. B. *et al.* Holocene climatic instability: a prominent, widespread event 8200 years ago. *Geology* **25**, 483–486 (1997).
- Rind, D. J. & Overpeck, J. T. Hypothesized causes of decade- to century-scale climatic variability: climate model results. *Quat. Sci. Rev.* **12**, 357–374 (1993).
- Shindell, D. T. *et al.* Solar forcing of regional climate change during the Maunder minimum. *Science* **294**, 2149–2152 (2001).
- Anderson, D. M., Overpeck, J. T. & Gupta, A. K. Increase in the Asian SW monsoon during the past four centuries. *Science* **297**, 596–599 (2002).
- Keigwin, L. D. The little ice age and medieval warm period in the Sargasso Sea. *Science* **274**, 1504–1508 (1996).
- Jennings, A. J. & Weiner, N. J. Environmental change in eastern Greenland during the last 1300 years: evidence from foraminifera and lithofacies in Nansen fjord, 68°N. *Holocene* **6**, 179–191 (1996).
- Manabe, S. & Stouffer, R. J. Century-scale effects of increased atmospheric CO₂ on the ocean-atmosphere system. *Nature* **364**, 215–218 (1993).
- Houghton, J. T. *et al.* (eds) *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, New York, 2001).
- Trenberth, K., Olson, J. & Large, W. A global ocean wind stress climatology based on ECMWF analyses. Technical Note NCAR/TN-388 + STR, 1–93 (National Center for Atmospheric Research, Colorado, Boulder, 1989).
- Reynolds, R. W. & Smith, T. M. A high-resolution global sea surface temperature climatology. *J. Clim.* **8**, 1571–1583 (1995).
- Janowiak, J. & Xie, P. CAMS-OP: A global satellite-rain gauge merged product for real-time precipitation monitoring applications. *J. Clim.* **12**, 3335–3342 (1999).
- Stuiver, M. & Reimer, P. J. Extended ¹⁴C data base and revised radiocarbon age calibration program. *Radiocarbon* **35**, 215–230 (1993).

Supplementary Information accompanies the paper on Nature’s website (<http://www.nature.com/nature>).

Acknowledgements A.K.G. thanks the US National Research Council for a senior research fellowship. J.T.O. thanks the US NSF Earth System History Program for supporting acquisition of AMS dates. We thank the Ocean Drilling Program for providing core samples.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.K.G. (e-mail: anilg@gg.iitkgp.ernet.in).

Reductive dehalogenation of chlorinated dioxins by an anaerobic bacterium

Michael Bunge*, Lorenz Adrian†, Angelika Kraus‡, Matthias Opel§, Wilhelm G. Lorenz‡, Jan R. Andreessen*, Helmut Görisch† & Ute Lechner*

* Institut für Mikrobiologie and ‡ Institut für Analytik und Umweltchemie, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle, Germany

† Fachgebiet Technische Biochemie, Institut für Biotechnologie der Technischen Universität Berlin, 13353 Berlin, Germany

§ GfA Gesellschaft für Arbeitsplatz- und Umweltanalytik mbH, 06896 Straach, Germany

Polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDDs and PCDFs) are among the most notorious environmental pollutants. Some congeners, particularly those with lateral chlorine substitutions at positions 2, 3, 7 and 8, are extremely toxic and carcinogenic to humans¹. One particularly promising mechanism for the detoxification of PCDDs and PCDFs is microbial reductive dechlorination. So far only a limited number of phylogenetically diverse anaerobic bacteria have been found that couple the reductive dehalogenation of chlorinated compounds—the substitution of a chlorine for a hydrogen atom—to energy conservation and growth in a process called dehalorespiration². Microbial dechlorination of PCDDs occurs in sediments and anaerobic mixed cultures from sediments, but the responsible organisms have not yet been identified or isolated. Here we show the presence of a *Dehalococcoides* species in four dioxin-dechlorinating enrichment cultures from a freshwater sediment highly contaminated with PCDDs and PCDFs. We also show that the previously described chlorobenzene-dehalorespiring bacterium *Dehalococcoides* sp. strain CBDB1 (ref. 3) is able to reductively dechlorinate selected dioxin congeners. Reductive dechlorination of 1,2,3,7,8-pentachlorodibenzo-*p*-dioxin (PeCDD) demonstrates that environmentally significant dioxins are attacked by this bacterium.

PCDDs and PCDFs are ubiquitous and recalcitrant environmental pollutants^{4,5}. Continuing anthropogenic contamination with PCDDs and PCDFs, formed as unwanted by-products of manufacturing and incineration processes, is of great public concern owing to the compounds' toxicity and tendency to bioaccumulate and biomagnify in wildlife and humans. Natural sources of dioxins include volcanic activities, forest fires, production by biological systems^{6,7} and as yet unknown formation processes^{8,9}. Because of their high hydrophobicity, dioxins are strongly adsorbed on organic matter and they therefore accumulate in aquatic sediments and soils, where conditions might be anaerobic. The only known biological process leading to a transformation of the highly chlorinated congeners under anaerobic conditions is the microbially mediated reductive dechlorination observed in microcosms or mixed cultures^{10–15}. Different sources of PCDDs and PCDFs introduce different complex mixtures of PCDD and PCDF congeners

into the environment. The extent to which intrinsic microbes change these source-specific profiles *in situ* is largely unknown, although studies of sediment cores of Lake Ketelmeer, a sedimentation area of the river Rhine in The Netherlands, have shown a change of congener distribution over time¹⁶. This observation can be taken as an indication that highly chlorinated dioxins are subject to anaerobic dehalogenation processes *in situ*. Our knowledge of the organisms involved in PCDD dechlorination is currently very limited. Until now, no pure culture with the ability to reductively dechlorinate dioxins has been described.

We have previously examined the reductive dehalogenation of selected dioxin congeners by anaerobic mixed cultures¹⁷. These enrichment cultures were established with various sediment samples from the stream Spittelwasser (Bitterfeld district, Germany), which contains dioxin at concentrations of up to 120,000 pg toxicity equivalents (I-TEQ) per g dry weight. Spiked 1,2,3,4-tetrachlorodibenzo-*p*-dioxin (TeCDD) (50 µM) was converted to a mixture of 1,3- and 2,3-dichlorodibenzo-*p*-dioxin (DiCDD). In previous experiments, the transformation pathways

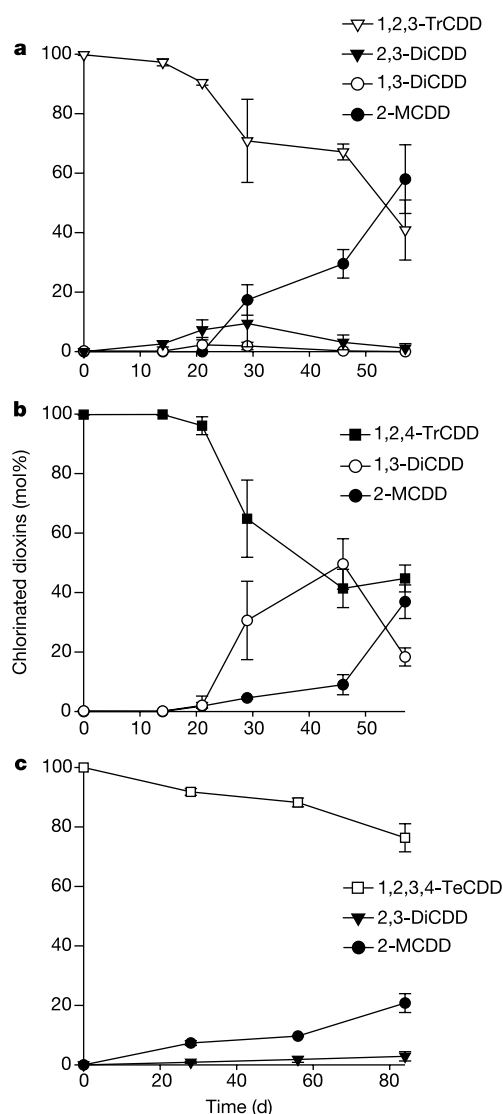


Figure 1 Time course of reductive dechlorination of 25 µM 1,2,3-TrCDD (a), 60 µM 1,2,4-TrCDD (b) and 46 µM 1,2,3,4-TeCDD (c) by *Dehalococcoides* sp. strain CBDB1. Molar distributions of the parent compounds and their dechlorination products are shown. Values are means and s.d. for triplicate samples. No dechlorination products were detected in sterile controls after 75 days (a, b) and 84 days (c).

were elucidated with subcultures spiked with the possible intermediate trichlorodibenzo-*p*-dioxins (TrCDDs) 1,2,4- and 1,2,3-TrCDD. A comparison of data obtained from the cultures with abiotic treatments (autoclaved and uninoculated controls) clearly demonstrated that the dechlorination was mediated by micro-organisms¹⁷.

In this study, the cultures from Spittelwasser sediment were transferred six times (10% v/v each time) into synthetic medium, to obtain sediment-free cultures reproducibly dechlorinating TrCDDs. 1,2,4-TrCDD was dechlorinated to 1,3-DiCDD, and 1,2,3-TrCDD was dechlorinated to 1,3-DiCDD and 2,3-DiCDD as observed in the initial cultures. However, in contrast to the initial cultures, most of the later transfers formed more 2,3- than 1,3-DiCDD from 1,2,3-TrCDD, demonstrating that dechlorination in the peripheral (*peri*)-position was now preferred. In addition, 2-monochlorodibenzo-*p*-dioxin (2-MCDD) was detected as the final dechlorination product of both TrCDDs. Most-probable-number analysis detected only about 10^4 dechlorinating bacteria ml^{-1} in the mixed cultures compared with a total cell number of more than 10^7 cells ml^{-1} . We therefore did not attempt isolation of the dioxin-dechlorinating bacterium by conventional agar-shake dilutions. Instead, a polymerase chain reaction (PCR)-based approach was used to study the presence of several bacteria with known dechlorination potential², among them *Dehalococcoides*. This genus comprises two strains with unusual dehalogenation properties: *Dehalococcoides* sp. strain CBDB1 (ref. 3) is the only known bacterium able to dechlorinate chlorinated benzenes, and *D. ethenogenes* strain 195 completely dechlorinates tetrachloroethene to ethene¹⁸. Using the oligonucleotide primers DET730 and DET1350 targeting the 16S ribosomal DNA (rDNA) of *Dehalococcoides*, PCR products were obtained from the sixth transfers of two 1,2,3- and two 1,2,4-TrCDD-dechlorinating enrichment cultures. Each of the four sequences had a length of about 600 base pairs (bp). They were identical with each other and also with the sequence of *Dehalococcoides* sp. strain CBDB1 (GenBank accession number AF230641), and shared 98.5% identity with the sequence of strain 195 (AF004928.2).

To substantiate an involvement of *Dehalococcoides* in dioxin dechlorination, the capability of strain CBDB1 to transform TrCDDs was studied. Liquid cultures pre-grown anaerobically on trichlorobenzenes were transferred (5% v/v inoculum) into a completely synthetic anaerobic medium³ containing 5 mM acetate as a carbon source, hydrogen as an electron donor as described³, and 25–60 μM 1,2,3- or 1,2,4-TrCDD. 1,2,3-TrCDD was dechlorinated to 1,3-DiCDD, 2,3-DiCDD and 2-MCDD, whereas 1,2,4-TrCDD was dechlorinated to 1,3-DiCDD and 2-MCDD (Fig. 1a, b). The time course of the dehalogenation of 1,2,3-TrCDD (Fig. 1a) showed 2,3-DiCDD as the initial and transient dechlorination product first detected after 14 days of incubation. Small amounts of 1,3-DiCDD and 2-MCDD were detectable after 21 days. Whereas concentrations of DiCDD remained at a low level, 2-MCDD concentrations steadily increased with decreasing levels of 1,2,3-TrCDD, leading to the transformation of about 60 mol% TrCDD to 2-MCDD within 57 days. In cultures to which 1,2,4-TrCDD had been added, the extent of reductive dechlorination was still low at day 21 (Fig. 1b). Although the fraction of 1,3-DiCDD increased rapidly thereafter, production of 2-MCDD was low during the first 46 days. At the end of the experiment (57 days), 37 mol% of TrCDD had been converted to 2-MCDD. Neither 1-MCDD nor non-substituted dibenzo-*p*-dioxin was detected throughout the study. Thus, 2-MCDD was the final dechlorination product of both TrCDDs. Autoclaved or uninoculated controls did not show any formation of dechlorinated products.

Three further dioxin congeners available to us were studied for reductive dechlorination by strain CBDB1. 2,3-DiCDD (10 μM) was transformed to 53 mol% 2-MCDD after 28 days of incubation. 1,2,3,4-TeCDD (46 μM) was dechlorinated within 84 days to 3 mol% 2,3-DiCDD and 21 mol% 2-MCDD (Fig. 1c). Traces of

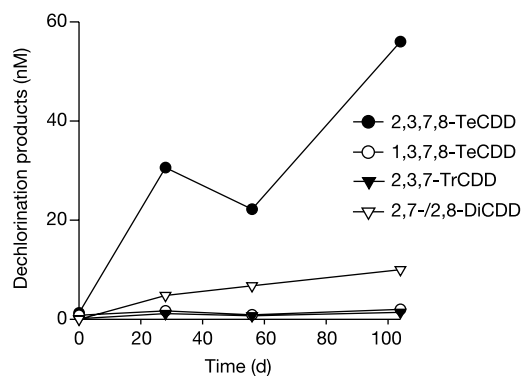


Figure 2 Formation of dechlorination products from 1,2,3,7,8-PeCDD. 1,2,3,7,8-PeCDD was added at a starting concentration of 3 μM . Within 104 days, 2.8 mol% was converted (decrease in PeCDD not shown) to products, which were formed in nanomolar concentrations. Values shown are means for two parallel samples.

1,3-DiCDD were detected only once (after 56 days). The concentrations of 1,2,3- and 1,2,4-TrCDD were below the detection limit throughout the experiment. 1,2,3,7,8-PeCDD (3 μM) was used as a model compound for dioxins chlorinated on both rings and as a representative of the 17 most toxic PCDD and PCDF congeners substituted at positions 2, 3, 7 and 8. The applied concentration (3 μM) was similar to the total PCDD and PCDF concentration (about 6 μM) determined in Spittelwasser sediment¹⁷. Pure cultures of strain CBDB1 transformed this compound, albeit slowly (2.8 mol% within 104 days), to 2,3,7,8-TeCDD, 2,7-DiCDD or 2,8-DiCDD (see Methods) and small amounts of 1,3,7,8-TeCDD and 2,3,7-TrCDD (Fig. 2). A control with autoclaved inoculum did not show any product formation from PeCDD.

To ensure that the reductive dechlorination of dioxins by strain CBDB1 was independent of the presence of chlorinated benzenes originating from the preculture, the cultures were subcultured sequentially (10% v/v each time) with either 1,2,3-TrCDD or with 1,2,4-TrCDD. The initial inoculum for this experiment originated from a culture spiked with 15 μM 1,2,3-trichlorobenzene and 15 μM 1,2,4-trichlorobenzene as the terminal electron acceptors. The cultures could be successfully transferred into synthetic medium with 1,2,3-TrCDD or 1,2,4-TrCDD at least four times (dilution factor 10,000). The dechlorination pattern remained the same through the four consecutive transfers. Adrian *et al.*³ previously showed that *Dehalococcoides* sp. strain CBDB1 does not grow in the synthetic medium without an added chlorinated electron acceptor. Because maintenance of the cultures was not dependent on the addition of chlorobenzenes, the data support the hypothesis that PCDD congeners are used as respiratory electron acceptors.

The dioxin dehalogenation reactions observed with strain CBDB1 are summarized in Fig. 3. 1,2,3,4-TeCDD was predominantly transformed by way of 1,2,3-TrCDD, as suggested by the formation of the subsequent intermediate 2,3-DiCDD. This indicates an initial dechlorination at *peri* positions. Dechlorination of spiked 1,2,3-TrCDD by CBDB1 also proceeded preferentially at the *peri* position. However, both DiCDDs are further dechlorinated either at a lateral or peripheral carbon to 2-MCDD as the final dechlorination product. Spiked 1,2,4-TrCDD was strictly *peri*-dechlorinated by the successive removal of chlorines in positions 1 and 4. Such a dechlorination pattern has also been observed in enrichment cultures from sediment layers from the Spittelwasser¹⁷ and the river Saale¹⁴. The shift from lateral to *peri* dechlorination of 1,2,3-TrCDD observed in the sixth transfers of the mixed cultures from Spittelwasser sediment suggests that *Dehalococcoides* sp. strain CBDB1 represented one subpopulation that was enriched by our culture conditions from a greater diversity of dioxin-dechlorinating bacteria in the sediment community.

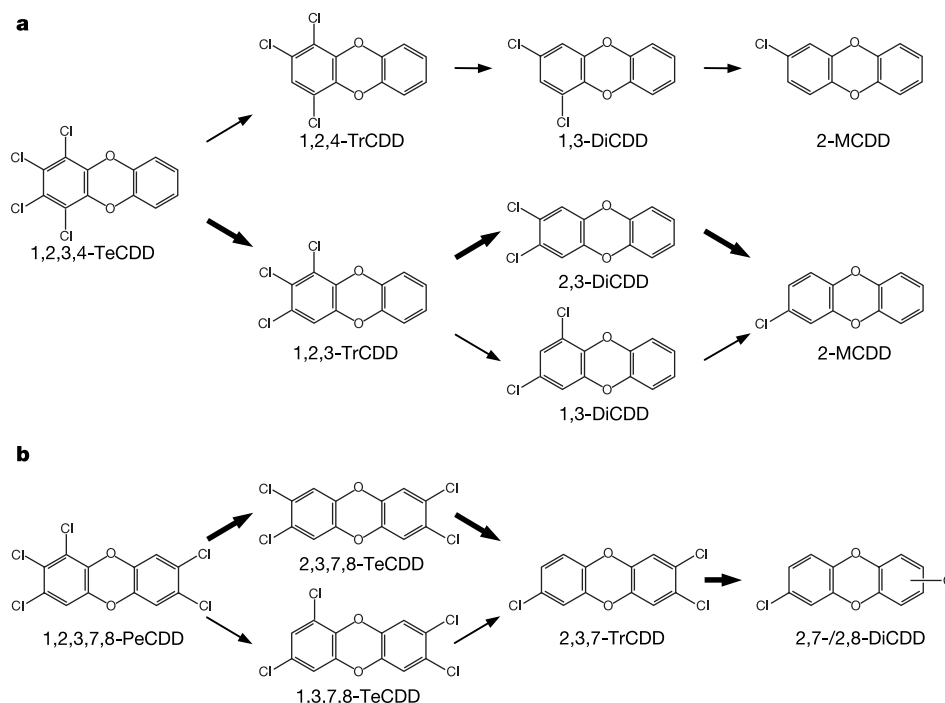


Figure 3 Proposed pathways of reductive dechlorination of spiked 1,2,3,4-TeCDD (**a**) and 1,2,3,7,8-PeCDD (**b**) by a pure culture of *Dehalococcoides* sp. strain CBDB1. The major routes are marked with bold arrows. The results of reductive dechlorination of spiked

1,2,4-TrCDD, 1,2,3-TrCDD and 2,3-DiCDD are included in **a**. For visualization purposes, identical structures of 1,3-DiCDD and 2-MCDD are shown inverted.

The removal of *peri*-chlorines from higher chlorinated 2,3,7,8-substituted dioxins involves the risk of forming 2,3,7,8-TeCDD, the most toxic dioxin congener (dioxin activation, upper pathway), as indicated earlier by other authors for sediment-derived mixed cultures¹³ and sediment microcosms¹⁵. However, these same cultures also exhibited dioxin detoxification (lower pathway) by the formation of less harmful TrCDDs and MCDD. This combined *peri*-lateral dechlorination pathway was assigned to non-methanogenic, non-spore-forming bacteria¹³.

We used 1,2,3,7,8-PeCDD, the immediate precursor of 2,3,7,8-TeCDD in the upper pathway, to study the dechlorination behaviour of strain CBDB1. It also produced 2,3,7,8-TeCDD but simultaneously the less toxic 1,3,7,8-TeCDD, 2,3,7-TrCDD and 2,7-/2,8-DiCDD. The removal of chlorine atoms from the individual rings of 1,2,3,7,8-PeCDD (as suggested by the products identified) strongly resembles the dechlorination patterns observed with 1,2,3-TrCDD and 2,3-DiCDD. This suggests a transient formation of 2,3,7,8-TeCDD and its further dechlorination through 2,3,7-TrCDD (Fig. 3). With strain CBDB1 it is therefore now possible to analyse and understand the dynamics of microbial transformation of environmentally significant dioxin congeners both in culture and in natural environments. This might help to predict a potential transient hazard involved in the overall detoxification process.

On the basis of 16S rDNA sequence, *Dehalococcoides* sp. strain CBDB1 is affiliated to a major subphylum of the phylum *Chloroflexi* (green non-sulphur bacteria). The tetrachloroethene-dehalorespiring bacterium *Dehalococcoides ethenogenes* strain 195 and *Dehalococcoides* sp. strain CBDB1 are so far the only isolated and cultivated representatives of the '*Dehalococcoides*' subphylum^{3,18}. For both strains, growth on fermentable substrates or the use of non-halogenated electron acceptors was not observed. Available 16S rDNA sequence information indicates the frequent presence of *Dehalococcoides*-related bacteria in mixed cultures that dehalogenate a variety of chlorinated substances^{19–22}. Additionally, two bacteria with distant relatedness to *Dehalococcoides* were described recently as being responsible for the dechlorination of poly-

chlorinated biphenyls^{23,24}. The high genetic potential of this genus for dechlorination reactions is evident from the genome sequence of *Dehalococcoides ethenogenes* (see the TIGR website at <http://www.tigr.org/tdb/mdb/mdbinprogress.html>). At least 15 reductive dehalogenase-homologue genes were detected in the genome, suggesting that different enzymes are involved in the reductive dehalogenation of different substrates. *Dehalococcoides* is thought to be well adapted for anaerobic reductive dehalogenation, which might be an ancient process for the turnover of naturally produced organohalogenes. Owing to their dehalogenation potential, indigenous and introduced organisms of the *Dehalococcoides* cluster are an important addition to the arsenal of organochlorine-transforming microorganisms² that are potentially applicable to the bioremediation of contaminated sites containing anthropogenic PCDD. □

Methods

Analytical techniques

MCDD, DiCDD and non-chlorinated dibenzo-*p*-dioxin were analysed from the headspace of the cultures by solid-phase microextraction (SPME, 100 μ m polydimethylsiloxane coated fibres; Supelco). The conventional method, including freeze-drying and concentration by evaporation¹⁴, yielded from the same samples only up to 21% of the 2-MCDD concentration compared with SPME, whereas the DiCDDs were found at similar concentrations. Samples were preconditioned for 2 h at 54 °C; fibres were equilibrated for 35 min at 54 °C and desorbed for 195 s (injection port: 260 °C) followed by splitless injection (0.7 min). The Shimadzu GC14 A/FID gas chromatograph was equipped with a DB-608 capillary column (30 m $\times$ 0.331 mm internal diameter, 0.5 μ m film thickness). A six-level external calibration curve from the headspace over minimal medium (amended with the respective dioxin concentrations ranging from 0.78 to 25 μ M) was generated. On the basis of the similar molar response within a given homologue group²⁵, dichlorinated dioxins were quantified with 2,7-DiCDD as a calibration standard. After SPME analysis, the liquid cultures were extracted and analysed for the respective tetrachlorinated and trichlorinated congeners as described¹⁴. Identification of monochlorodioxins and dichlorodioxins was confirmed on a gas chromatograph with mass-selective detector (GC/MSD; HP6890/HP5973) from mass spectra and the relative retention of authentic standards¹⁴.

For every data point, two or three individual 3-ml cultures were harvested. Heterogeneities of absolute dioxin concentrations, which might have been due to slight differences in starting concentrations, recovery efficiencies and sorption onto differing cell numbers, were normalized by the expression of each congener as mol% of the sum of all congeners detected. 1,2,3,7,8-PeCDD and its dechlorination products were analysed on a

GC/MSD after extraction of the 3-ml cultures with toluene, gentle concentration and separation on a SP-2331 column (60 m × 0.25 mm internal diameter, 0.2 µm film thickness) in accordance with standard conditions for dioxin analyses²⁶. The relative retention times and response factors were determined by analysing calibration mixtures containing five native congeners (1,2,3,7,8-PeCDD, 2,3,7,8-TeCDD, 1,2,4-TrCDD, 2,7-DiCDD and 1-MCDD) and five ¹³C₁₂-labelled internal standards (1,2,3,7,8-PeCDD, 2,3,7,8-TeCDD, 1,3,6,8-TeCDD, 2,3,7-TrCDD and 2,8-DiCDF). Before extraction, the five ¹³C₁₂-labelled congeners were added and the recovery efficiency compared with that of ¹³C₆-1,2,3,4-TeCDD was determined. The recovery efficiency ranged between 75% and 100%. 2,3-DiCDD and 2,7-DiCDD could be identified by relative retention compared with ¹³C₁₂-2,3,7-TrCDD, but 2,7-DiCDD and 2,8-DiCDD were not expected to be separated under these conditions.

Cultivation

Chlorinated dibenzo-*p*-dioxins (amchro, Hattersheim, Germany) were dissolved in acetone and added to each cultivation tube; the solvent was evaporated by using 20% CO₂/80% N₂. The synthetic culture medium³ was prepared with strict anaerobic techniques. *Dehalococcoides* sp. strain CBDB1 was cultivated under anaerobic conditions with a gas phase of 20% CO₂/80% N₂ in several parallel cultures in 15-ml Hungate tubes (Bellco Glass, Inc., Vineland, New Jersey, USA) containing 3 ml of culture volume sealed with thick butyl rubber stoppers (Ochs Glasgerätebau, Bovenden, Germany), which were best suited to the maintenance of highly reduced conditions over long cultivation times, essential for the growth of strain CBDB1. For some experiments Teflon disks were placed below the septa to reduce potential sorption on the stoppers. Recovery of total dioxins decreased over time. After 4 weeks of incubation, recovery ranged between 54% and 92% without Teflon coats, and between 38% and 76% with the coats. The cultures were supplied with 5 mM acetate and hydrogen (2.5 ml injected into the headspace of the Hungate tubes, corresponding to an approximate dissolved hydrogen concentration of 0.1 mM). The consecutive transfers into fresh medium with acetate, hydrogen and TrCDDs were performed every 24 days. The cultures were incubated at 30 °C in the dark and agitated at 115 r.p.m. A comparison with non-agitated cultures demonstrated that agitation promoted dechlorination. For example, after 56 days, 60 mol% of the 1,2,3-TrCDD was converted in an agitated culture and only 40 mol% in a static culture. Controls were established with autoclaved inocula, and without inoculum.

PCR detection and sequence analysis

Community DNA, extracted from dioxin-dechlorinating enrichment cultures with the use of standard methods, was used as a template for almost complete 16S rDNA amplification with the universal primer pair fD1 and rP2. PCR amplification and DNA sequencing methods have been described in detail previously²⁷. Amplified products from the initial PCR were then used as templates in the second amplification with *Dehalococcoides*-targeted primers DET730 (5'-GCGGTTTCTAGGTTGTC-3') and DET1350 (5'-CACCTTGCTGATATGCGG-3'), which were specifically designed with ARB software²⁸. Obtained amplicons were sequenced entirely in both directions and analysed as described²⁷.

Received 29 July; accepted 11 October 2002; doi:10.1038/nature01237.

- Kaiser, J. Just how bad is dioxin? *Science* **288**, 1941–1944 (2000).
- Holliger, C., Wohlfarth, G. & Diekert, G. Reductive dechlorination in the energy metabolism of anaerobic bacteria. *FEMS Microbiol. Rev.* **22**, 383–398 (1999).
- Adrian, L., Szezyk, U., Wecke, J. & Görsch, H. Bacterial dehalorespiration with chlorinated benzenes. *Nature* **408**, 580–583 (2000).
- Brzuz, L. P. & Hites, R. A. Global mass balance for polychlorinated dibenzo-*p*-dioxins and dibenzofurans. *Environ. Sci. Technol.* **30**, 1797–1804 (1996).
- Meharg, A. A. & Osborn, D. Dioxins released from chemical accidents. *Nature* **375**, 353–354 (1995).
- Gribble, G. W. *Encyclopedia of Environmental Analysis and Remediation* (ed. Meyers, R. A.) 972–1035 (Wiley, New York, 1998).
- Hoekstra, E. J., de Weerd, H., de Leer, E. W. B. & Brinkman, U. A. T. Natural formation of chlorinated phenols, dibenzo-*p*-dioxins, and dibenzofurans in soil of a Douglas fir forest. *Environ. Sci. Technol.* **33**, 2543–2549 (1999).
- Rappe, C. *et al.* PCDDs in naturally-formed lake sediment cores from Southern Mississippi, USA. *Organohalogen Comp.* **43**, 111–116 (1999).
- Müller, J. F. *et al.* PCDDs, PCDFs, PCBs and HCB in marine and estuarine sediments from Queensland, Australia. *Chemosphere* **39**, 1707–1721 (1999).
- Adriaens, P. & Grbic-Galic, D. Reductive dechlorination of PCDD/F by anaerobic cultures and sediments. *Chemosphere* **29**, 2253–2259 (1994).
- Adriaens, P., Fu, Q. & Grbic-Galic, D. Bioavailability and transformation of highly chlorinated dibenzo-*p*-dioxins and dibenzofurans in anaerobic soils and sediments. *Environ. Sci. Technol.* **29**, 2252–2260 (1995).
- Beurskens, J. E. M. *et al.* Dehalogenation of chlorinated dioxins by an anaerobic microbial consortium from sediment. *Environ. Toxicol. Chem.* **14**, 939–943 (1995).
- Barkovskii, A. L. & Adriaens, P. Microbial dechlorination of historically present and freshly spiked chlorinated dioxins and diversity of dioxin dechlorinating populations. *Appl. Environ. Microbiol.* **62**, 4556–4562 (1996).
- Ballerstedt, H., Kraus, A. & Lechner, U. Reductive dechlorination of 1,2,3,4-tetrachlorodibenzo-*p*-dioxin and its products by anaerobic mixed cultures from Saale River sediment. *Environ. Sci. Technol.* **31**, 1749–1753 (1997).
- Albrecht, I. D., Barkovskii, A. L. & Adriaens, P. Production and dechlorination of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin in historically-contaminated estuarine sediments. *Environ. Sci. Technol.* **33**, 737–744 (1999).
- Beurskens, J. E. M., Mol, G. A. J., Barreveld, H. L., van Munster, B. & Winkels, H. J. Geochronology of priority pollutants in a sedimentation area of the Rhine River. *Environ. Toxicol. Chem.* **12**, 1549–1566 (1993).

- Bunge, M., Ballerstedt, H. & Lechner, U. Regiospecific dechlorination of spiked tetra- and trichlorodibenzo-*p*-dioxins by anaerobic bacteria from PCDD/F-contaminated Spittelwasser sediments. *Chemosphere* **43**, 675–681 (2001).
- Maymó-Gatell, X., Chien, Y., Gossett, J. M. & Zinder, S. H. Isolation of a bacterium that reductively dechlorinates tetrachloroethene to ethene. *Science* **276**, 1568–1571 (1997).
- Dojka, M. A., Hugenoltz, P., Haack, S. K. & Pace, N. R. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64**, 3869–3877 (1998).
- Hendrickson, E. R. *et al.* Molecular analysis of *Dehalococcoides* 16S ribosomal DNA from chloroethene-contaminated sites throughout North America and Europe. *Appl. Environ. Microbiol.* **68**, 485–495 (2002).
- von Wintzingerode, F., Selent, B., Hegemann, W. & Göbel, U. B. Phylogenetic analysis of an anaerobic, trichlorobenzene-transforming microbial consortium. *Appl. Environ. Microbiol.* **65**, 283–286 (1999).
- Löffler, F. E., Sun, Q., Li, J. & Tiedje, J. M. 16S rRNA gene-based detection of tetrachloroethene-dechlorinating *Desulfuromonas* and *Dehalococcoides* species. *Appl. Environ. Microbiol.* **66**, 1369–1374 (2000).
- Wu, Q., Watts, J. E. M., Sowers, K. R. & May, H. D. Identification of a bacterium that specifically catalyzes the reductive dechlorination of polychlorinated biphenyls with doubly flanked chlorines. *Appl. Environ. Microbiol.* **68**, 807–812 (2002).
- Cutter, L. A., Watts, J. E. M., Sowers, K. R. & May, H. D. Identification of a microorganism that links its growth to the reductive dechlorination of 2,3,5,6-chlorobiphenyl. *Environ. Microbiol.* **3**, 699–709 (2001).
- Ballschmiter, K., Zoller, W., Schäfer, W. & Class, T. Quantitation of polychlorodibenzodioxin and polychlorobiphenyl standards by gas-chromatography-flame ionisation detection. *Fresenius Z. Anal. Chem.* **321**, 247–251 (1985).
- DIN EN 1948-3 Stationary source emissions—determination of the mass concentration of PCDDs/PCDFs. Part 3. Identification and quantification (Beuth, Berlin, 1997).
- Breitenstein, A., Saano, A., Salkinoja-Salonen, M., Andreesen, J. R. & Lechner, U. Analysis of a 2,4,6-trichlorophenol-dehalogenating enrichment culture and isolation of the dehalogenating member *Desulfotobacterium fropieri* strain TCP-A. *Arch. Microbiol.* **175**, 133–142 (2001).
- Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568 (1998).

Acknowledgements We thank R. Weber for providing standards for 1,2-, 1,3- and 1,4-DiCDD; M. M. Häggblom and R. U. Halden for comments on the manuscript; and H. Ballerstedt for helpful discussion and support. This work was supported by grants of the Land Sachsen-Anhalt and the Deutsche Forschungsgemeinschaft to U.L. and M.B., of the Technische Universität Berlin to L.A. and of the Fonds der Chemischen Industrie to J.R.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.B. (e-mail: m.bunge@mikrobiologie.uni-halle.de).

Directional postcopulatory sexual selection revealed by artificial insemination

Jonathan P. Evans, Lorenzo Zane, Samuela Francescato & Andrea Pilastro

Department of Biology, University of Padova, via U. Bassi 58/B, Padova, 1-35131, Italy

Postcopulatory sexual selection comprises both sperm competition, where the sperm from different males compete for fertilization¹, and cryptic female choice, where females bias sperm use in favour of particular males². Despite intense current interest in both processes as potential agents of directional sexual selection³, few studies have attributed the success of attractive males to events that occur exclusively after insemination. This is because the interactions between pre- and post-insemination episodes of sexual selection can be important sources of variation in paternity⁴. The use of artificial insemination overcomes this difficulty because it controls for variation in male fertilization success attributable to the female's perception of male quality, as well as effects due to mating order and the relative contribution of sperm from competing males⁵. Here, we adopt this technique and show that in guppies, when equal numbers of sperm from two males compete for fertilization, relatively colourful individuals

achieve greater parentage than their less ornamented counterparts. This finding indicates that precopulatory female mating preferences can be reinforced exclusively through postcopulatory processes occurring at a physiological level. Our analysis also revealed that relatively small individuals were advantaged in sperm competition, suggesting a possible trade-off between sperm competitive ability and body growth.

Postcopulatory sexual selection, arising from polyandry, is a pervasive evolutionary force that permits directional selection in the form of sperm competition¹ and cryptic female choice^{2,3}. In experimental studies, the outcome of postcopulatory sexual selection is traditionally described in terms of the proportion of offspring sired by the second male to copulate with a female, formally defined as the P_2 value⁴. However, values of P_2 are typically characterized by extreme and often unexplained variation⁶ which can obscure the relative importance of pre- and post-insemination processes of sexual selection. The use of artificial insemination overcomes this difficulty because it uncouples precopulatory sexual selection (notably mate choice) from postcopulatory selective processes. The technique therefore allows us to distinguish between the fertilizing capacities of different males⁷ and to test explicitly whether directional sexual selection can proceed exclusively through postcopulatory processes at a physiological level⁵. Here we use artificial insemination to examine the importance of postcopulatory sexual selection on preferred male traits in the guppy *Poecilia reticulata*, an internally fertilizing species of freshwater fish. Recent evidence reveals that attractive male guppies (those that were accepted as mates more quickly) achieve greater parentage after natural double matings⁸. Our aim is to determine whether the success of attractive males can be attributed to events occurring exclusively during post-insemination episodes of selection, as predicted by a positive association between male sexual attractiveness and sperm quality^{9–11}.

Guppies have a promiscuous, non-resource-based mating system¹² that includes both female choice and sexual coercion by males¹³. During precopulatory mate choice, females typically prefer relatively colourful males exhibiting high rates of courtship¹³. In particular, the area of carotenoid coloration in males (including orange, red and yellow) consistently influences female mating decisions^{13,14}. We therefore specifically focused our analysis of

male phenotype on body coloration and mating behaviour, but also included body size because this trait has been shown to influence female choice in some¹⁵, but not all studies (see ref. 14). In each trial, the relative phenotype scores of two males, arbitrarily labelled A and B, were related to each male's subsequent share of paternity following the artificial insemination of equal-sized ejaculates from both males. For each of the broods we calculated the proportion of offspring sired by the focal male B (hereafter termed P_B) and related this parameter to differences in the behavioural and morphological phenotypes of the two putative sires. Thirty-five inseminations were performed yielding $n = 27$ broods (mean number of offspring per brood, 11.1 ± 5.5 s.d.; range 3–25), which is a similar rate of success to that observed after natural double matings in this population⁸. Mean body size and the extent of body coloration (mean trait measurements for the two males in each trial) did not differ significantly between males involved in unsuccessful ($n = 8$) and successful trials ($n = 27$) (Student's t -test: body size $t_{8,27} = 0.39$, $P = 0.70$; total body coloration $t_{8,27} = 1.71$, $P = 0.11$; all individual colour components not significant).

The resultant paternity distribution ranged from 0 to 100% (mean, 0.53 ± 0.33 s.d.). The variance in P_B was significantly larger than the expected binomial variance (James test statistic, 15.1, $P < 0.0001$)¹⁶, which is assumed under the null hypothesis that sperm from each male have equal chances of fertilizing ova—the so-called 'fair raffle' model of sperm competition¹ (Fig. 1). Furthermore, a logistic regression analysis revealed that differences in the phenotype of competing males accounted for significant deviance in P_B (Table 1). Within pairs of males, individuals with relatively more orange sired a greater proportion of offspring than predicted by chance, whereas the area of black and blue did not predict male fertilization success (Fig. 2a–c). We suggest that this finding is unlikely to be explained by the differential survival of embryos from attractive and unattractive males because such an interpretation requires that broods sired exclusively by the more colourful male contain more offspring than those with shared paternity (and that those with shared paternity contain more offspring than families sired entirely by the drab male). Our results revealed that brood size was not correlated with the proportion of offspring sired by the most colourful male in each pair ($r = 0.23$, $P = 0.26$, $n = 27$) and

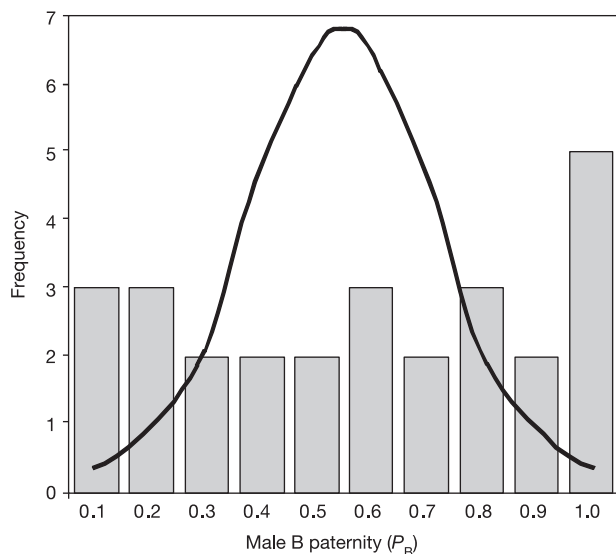


Figure 1 Expected versus observed paternity distribution. The expected paternity distribution, based on the assumption of random sperm use (line), is compared with the observed values of P_B (bars). The expected P_B curve is the mean of 10,000 randomized distributions (Monte Carlo simulation), where both putative sires per family were assumed to have an equal chance of siring each individual offspring.

Table 1 Proportion of offspring sired in relation to male phenotype

Generalized linear model with binomial errors*					
	d.f.	Deviance	Mean deviance	Ratio	P
Regression	2	7.84	3.918	3.92	0.02
Residual	24	29.23	1.218		
Total	26	37.07	1.426		
Parameters included in the model					
	Estimate	s.e.m.	t	P	
Constant	0.372	0.278	1.34	0.180	
Difference in orange†	0.089	0.040	2.23	0.026	
Difference in body size	−0.430	0.177	−2.43	0.015	
Parameters excluded from the model					
	t	P			
Difference in sigmoid rate	0.85	ns			
Difference in black	0.67	ns			
Difference in blue	1.20	ns			

*Final model (logit transformation): response variate, offspring sired by male B; binomial totals, brood size; fitted terms, difference in absolute orange coloration and difference in body size. Overdispersion was corrected using the Williams procedure³⁰ ($\phi = 0.306$).

†We obtained the same results when the three colour components were combined using principal components analysis: two components were extracted from the original variables (difference in colour area); the first component (PC1) explained 62.7% of the variance and was correlated with the differences in orange ($r = 0.962$), the second component (PC2) explained 25.3% of the variance and was negatively correlated with blue ($r = -0.956$) and positively with black ($r = 0.261$). Once entered in the generalized linear model, only the first component was a significant predictor of paternity share (PC1: $t = 2.48$, $P = 0.013$, difference in body length: $t = -2.54$, $P = 0.011$; PC2: $t = -0.83$, $P = 0.41$). Qualitatively similar results were obtained when percentage, rather than absolute colour was used (analysis not shown).

d.f., degrees of freedom.

furthermore that brood size was not a function of the mean extension of orange of the two males in each trial ($r = 0.15$, $P = 0.44$, $n = 27$).

We also found that body size was related to male success, with relatively small males achieving greater paternity (Fig. 2d). Body size was not significantly correlated with either absolute or relative orange body coloration (data pooled to include all $n = 54$ males; $r = 0.18$, $P = 0.21$; $r = -0.04$, $P = 0.78$, respectively). Again, there was no evidence that the success of relatively small males was due to the enhanced survival of their embryos or offspring. Indeed, brood size tended to be positively, rather than negatively associated with the mean body size of the two males in each trial (mean body length, $r = 0.35$, $P = 0.07$, $n = 27$), although this trend was less apparent when one family, a statistical outlier with respect to the others, was excluded from the analysis ($r = 0.27$, $P = 0.19$, $n = 26$). Although the resorption of embryos has never been documented in poeciliids¹⁷, further work is required to determine the extent of embryo mortality in guppies, and furthermore whether this phenomenon is non-random with respect to male phenotype. Until this information is obtained, the possibility that the variation in brood size observed in this study, and its possible influence on paternity share, was due to natural selection for enhanced embryo survival cannot be fully discounted. Finally, male courtship behaviour, estimated for each male before the artificial insemination trials, did not predict male B paternity (Table 1).

To our knowledge, these results provide the first evidence that when behavioural interactions between males and females are prevented, postcopulatory selection can favour the same traits that are preferred by females during precopulatory mate choice. The colour components measured in this study, and in particular the relative area of orange, consistently predict female mating

preferences in guppies (see, for example, refs 14, 18, 19), including the population chosen for this study (A.P., A. Bisazza and J.P.E., unpublished data). In previous studies, such congruence between pre- and postcopulatory sexual selection has been reported but may have arisen because attractive males inseminate more sperm²⁰, or because female behaviour can favour particular males in sperm competition^{2,21–23}. In our experiment, however, ejaculate size was controlled and females were denied the opportunity of assessing male quality, and therefore such effects cannot explain the finding that relatively colourful males were more successful. In addition, our results are unlikely to be explained by female sperm selection because this mechanism requires that the haploid expression of sperm reflect male attractiveness. Although haploid gene expression does occur in nature, the examples to date are confined to cases involving compatibility-based discrimination against particular sperm genotypes rather than directional postcopulatory sexual selection (see refs 3, 24 and references therein). Such compatibility effects, in addition to other non-directional selective processes (for example, poor sperm mixing²⁵) cannot account for the success of colourful males, although it is possible that both factors contributed towards the observed variation in P_B in this study.

Our results suggest instead that males exhibiting high levels of coloration produce competitively superior ejaculates⁹. Whether this effect is due to condition-dependence (see below) or the enhanced genetic quality of colourful males remains to be investigated. However, our results also reveal that relatively small males achieve greater parentage than larger individuals. Unlike colour, the influence of male body size on female preferences is less clear, with some studies reporting female preferences for large males¹⁵ and others revealing no discrimination by females for this trait¹⁴. Irrespective of whether body size is under selection through precopulatory female choice, for the reasons given above it seems unlikely that sperm choice by females accounted for the small male advantage uncovered by our study. Thus, to the extent that sperm competition, rather than sperm choice, accounts for our findings, the two phenotypic traits identified in this study (body size and coloration) appear to covary with sperm competitive ability in opposing directions. Such effects would be predicted, for example, if sperm quality were to depend on an adequate intake of carotenoids in the diet¹⁰ as well as energy traded-off against body growth²⁶. An important direction for future research is to identify the specific sperm traits that confer reproductive advantages on males and examine how they interact with the phenotypic traits identified in our study. □

Methods

Origin of fish and assessment of male mating behaviour

Guppies, which were first-generation descendants of wild-caught fish from the Tacarigua River, Trinidad, were maintained as previously described⁸. Behavioural trials took place between 09.00 and 14.00. On the evening before each trial an adult, non-virgin female and two sexually immature juvenile fish were placed into an observation tank (46 cm × 19 cm × 27 cm). On the following morning a test male (aged six months), unfamiliar to the female and juvenile fish, was placed in the tank to settle for 30 min. The test males were fully mature and were selected only if they were in good condition and displayed sexual behaviour (as evident from their behaviour in the stock tank before capture). Juveniles were included to facilitate the settlement of the adult fish. The rate of male courtship (sigmoid displays) was measured over a 10-min period following an established protocol²⁷. After each trial the female and juveniles were replaced by different fish for the following day's trial. The males were isolated for 3 days before the phenotype measurements and the extraction of sperm for artificial insemination.

Male ornamentation

Test males were paired at random, anaesthetized (MS222) and photographed with a digital camera. Image analysis software (Scion Corporation, <http://www.scioncorp.com>) was used to measure total length and analyse body coloration. Briefly, three main components of these colour patterns were considered: the area of (1) carotenoid pigmentation (including orange, red and yellow), (2) melanistic black spots, and (3) the iridescent structural colours, which include blue and green. These three components are termed orange, black and blue respectively in the main text. There were no significant differences between paired males (labelled A and B) in body size (paired t -test; $t_{26} = 0.75$, $P = 0.46$), courtship behaviour ($t_{23} = 0.01$, $P = 0.99$; behaviour not obtained from $n = 3$ males) or

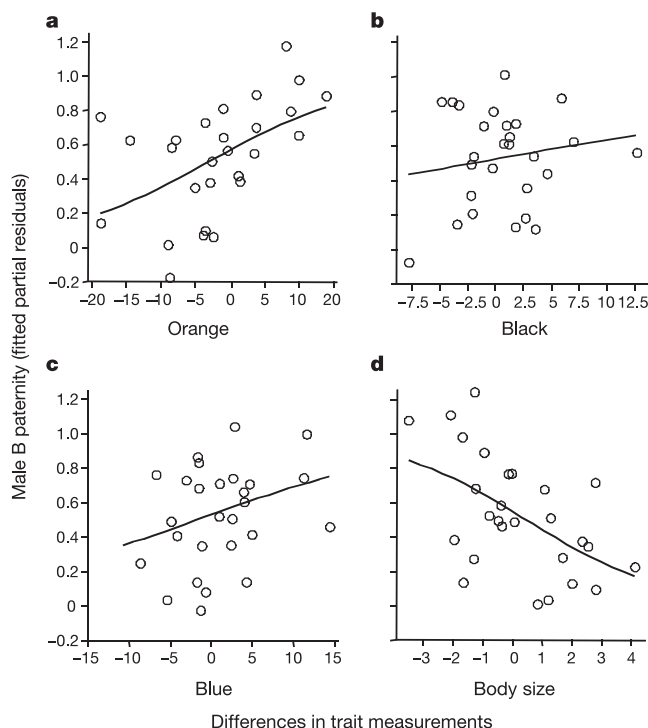


Figure 2 Effect of phenotypic traits on the proportion of offspring sired. The relationship between the proportion of offspring sired by male B (P_B) and the differences in phenotype scores (male B trait minus male A trait) between competing males. Differences in **a**, orange, **b**, black and **c**, blue areas of colour spots (mm^2), are plotted. **d**, Difference in body size (mm) is plotted. Points correspond to the partial residuals of the observed data with respect to the fitted curves (see Table 1).

overall body coloration ($t_{26} = 0.48$, $P = 0.63$; all individual components not significant). The analysis of body coloration was done blind of paternity assignment (see below).

Artificial insemination

Male guppies produce sperm packaged in bundles (spermatozeugmata). In preliminary trials, 23 males were repeatedly stripped to obtain three independent measures for the number of sperm per bundle for each male (for two males only two counts were obtained). Analysis of variance (with males as random factors) revealed that between-male variation in the number of sperm per bundle did not exceed that observed within individuals (analysis of variance (ANOVA): $F_{22,44} = 0.76$, $P = 0.76$). The number of sperm per bundle was not significantly correlated with any of the morphological traits measured in this study (orange: $r = 0.20$, $P = 0.35$; blue: $r = 0.14$, $P = 0.52$; black: $r = 0.03$, $P = 0.91$; body size: $r = 0.03$, $P = 0.88$, $n = 23$). In each trial, equal numbers of bundles were obtained from each of the two males²⁷, gently mixed and inseminated simultaneously into an anaesthetized female using a machine-pulled micropipette (penetration depth approximately 2 mm). The number of sperm inseminated was based on the size of natural ejaculates in the study population ($\sim 0.5 \times 10^6$ spermatozoa) (see also ref. 20). After artificial insemination, females were revived and isolated until they produced their first brood. Tissue samples were obtained from all fish (mother, two putative sires and offspring) for the subsequent paternity analysis.

Paternity analysis

We observed no offspring mortality before paternity assignment. Two microsatellite markers (accession numbers: AF164205 and AF026459) were used to estimate each male's relative share of paternity (P_B). The polymerase chain reaction (PCR) protocol followed previous work⁸ with the exception that one primer from each pair was end-labelled with a fluorescent phosphoramidite dye. Amplified fragments were separated by electrophoresis on an ABI 3100 sequencer (ABI PRISM), using 400 HD ROX (Perkin-Elmer) as a size standard. PCR products were visualized using Genographer software (v. 1.6.0) and paternity was assigned to offspring according to allele sharing between putative sires, mother and offspring. Both loci were highly variable (18 and 33 alleles, with expected heterozygosities of 0.901 and 0.941 respectively) and exhibited a global exclusion probability of 0.92. Where one parent (the mother) is known, paternity assignment with two putative sires (as in our study) is calculated to be 100% (Cervus v. 2.0)²⁸. In practice, paternity was unambiguously assigned to all offspring ($n = 300$ from 27 broods).

Data analysis

The James method¹⁶ was used to test whether the proportion of offspring sired by the male labelled B fluctuates randomly between families (with different brood sizes)²⁹, and thus whether the observed P_B variance deviates from binomial expectation. A generalized linear model with binomial errors and logit link function was then used to determine whether male phenotype accounted for deviance in P_B . For each family, male B success (the number of offspring sired by male B) and the total number of offspring were entered as the binomial response variables. Predictor variables, representing differences in the phenotypic trait measurements taken from the two males per family (male B trait minus male A trait), were fitted into the model. Initially, the full model included all possible explanatory variables; the term with the least significant probability was then excluded in a stepwise procedure. The change in deviance in the generalized linear model resulting from the removal of each term was tested against an F -distribution. We removed all terms whose exclusion did not cause a significant change in the deviation of the model (Table 1).

Received 18 October; accepted 10 December 2002; doi:10.1038/nature01367.

- Parker, G. A. in *Sperm Competition and Sexual Selection* (eds Birkhead, T. R. & Møller, A. P.) 3–54 (Academic, San Diego, 1998).
- Eberhard, W. G. *Female Control: Sexual Selection by Cryptic Female Choice* (Princeton Univ. Press, Princeton, 1996).
- Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3**, 262–273 (2002).
- Birkhead, T. R. & Møller, A. P. (eds) *Sperm Competition and Sexual Selection* (Academic, San Diego, 1998).
- Birkhead, T. R. Defining and demonstrating postcopulatory female choice—again. *Evolution* **54**, 1057–1060 (2000).
- Simmons, L. W. *Sperm Competition and its Evolutionary Consequences in the Insects* (Princeton Univ. Press, Princeton, New Jersey, 2001).
- Dziuk, P. J. Factors that influence the proportion of offspring sired by a male following heterospermic insemination. *Anim. Reprod. Sci.* **43**, 65–68 (1996).
- Evans, J. P. & Magurran, A. E. Patterns of sperm precedence and predictors of paternity in the Trinidadian guppy. *Proc. R. Soc. Lond. B* **268**, 719–724 (2001).
- Sheldon, B. C. Male phenotype, fertility, and the pursuit of extra-pair copulations by female birds. *Proc. R. Soc. Lond. B* **257**, 25–30 (1994).
- Blount, J. D., Møller, A. P. & Houston, D. C. Antioxidants, showy males and sperm quality. *Ecol. Lett.* **4**, 393–396 (2001).
- Jennions, M. D. & Petrie, M. Why do females mate multiply? A review of the genetic benefits. *Biol. Rev.* **75**, 21–64 (2000).
- Evans, J. P. & Magurran, A. E. Multiple benefits of multiple mating in guppies. *Proc. Natl Acad. Sci. USA* **97**, 10074–10076 (2000).
- Houde, A. E. *Sex, Color, and Male Choice in Guppies* (Princeton Univ. Press, Princeton, New Jersey, 1997).
- Endler, J. A. & Houde, A. E. Geographical variation in female preferences for male traits in *Poecilia reticulata*. *Evolution* **49**, 456–468 (1995).
- Reynolds, J. D. & Gross, M. R. Female mate preference enhances offspring growth and reproduction in a fish. *Poecilia reticulata*. *Proc. R. Soc. Lond. B* **264**, 57–62 (1992).
- James, W. H. The distribution of the combinations of the sexes in mammalian litters. *Genet. Res.* **26**, 45–53 (1975).

- Constantz, G. D. in *Ecology and Evolution of Livebearing Fishes* (eds Meffe, G. K. & Snelson, Jr F. F.) 33–50 (Prentice-Hall, New Jersey, 1989).
- Endler, J. A. Natural and sexual selection on colour patterns in poeciliid fishes. *Environ. Biol. Fishes* **9**, 173–190 (1983).
- Houde, A. E. Mate choice based on naturally occurring colour-pattern variation in a guppy population. *Evolution* **41**, 1–10 (1987).
- Pilastro, A., Evans, J. P., Sartorelli, S. & Bisazza, A. Male phenotype predicts insemination success in guppies. *Proc. R. Soc. Lond. B* **269**, 1325–1330 (2002).
- Pizzari, T. & Birkhead, T. R. Female feral fowl eject sperm of subdominant males. *Nature* **405**, 787–789 (2000).
- Lewis, S. M. & Austad, S. N. Sexual selection in four beetles: the relationship between sperm precedence and male olfactory attractiveness. *Behav. Ecol.* **5**, 219–224 (1994).
- Edvardsson, M. & Arnqvist, G. Copulatory courtship and cryptic female choice in red flour beetles *Tribolium castaneum*. *Proc. R. Soc. Lond. B* **267**, 559–563 (2000).
- Zeh, J. A. & Zeh, D. W. The evolution of polyandry II: post-copulatory defences against genetic incompatibility. *Proc. R. Soc. Lond. B* **264**, 69–75 (1997).
- Harvey, I. F. & Parker, G. A. 'Sloppy' sperm mixing and intraspecific variation in sperm precedence (P_2) patterns. *Proc. R. Soc. Lond. B* **267**, 2537–2542 (2000).
- Neems, R. M., Lazarus, J. & McLachlan, A. J. Lifetime reproductive success in a swarming midge: trade-offs and stabilizing selection for male body size. *Behav. Ecol.* **9**, 279–286 (1998).
- Matthews, I. M., Evans, J. P. & Magurran, A. E. Male display rate reveals ejaculate characteristics in the Trinidadian guppy *Poecilia reticulata*. *Proc. R. Soc. Lond. B* **264**, 695–700 (1997).
- Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7**, 639–655 (1998).
- Krackow, S., Meelis, F. & Hardy, I. C. W. in *Sex Ratios: Concepts and Research Methods* (ed. Hardy, I. C. W.) 112–131 (Cambridge Univ. Press, Cambridge, UK, 2002).
- Williams, D. A. Extra-binomial variation in logistic linear models. *Appl. Stat.* **31**, 144–148 (1982).

Acknowledgements We thank T. Birkhead, J. Kelley, B. Kempenaers, F. Neat, T. Pizzari, L. Simmons and A. Skinner for comments on an earlier draft of the manuscript, C. Romualdi for statistical advice and A. Ludlow for assistance with the preliminary artificial insemination trials. This research was supported by a Marie Curie Independent Research Fellowship from the EU and was carried out in conformity with the relevant Italian laws governing the care of animals in research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P.E. (e-mail: jonathan.evans@unipd.it).

Comparative power curves in bird flight

B. W. Tobalske*, T. L. Hedrick†, K. P. Dial‡ & A. A. Biewener†

* Department of Biology, University of Portland, 5000 North Willamette Boulevard, Portland, Oregon 97203, USA

† Concord Field Station, Harvard University, Old Causeway Road, Bedford, Massachusetts 01738, USA

‡ Division of Biological Sciences, University of Montana, Missoula, Montana 59812, USA

The relationship between mechanical power output and forward velocity in bird flight is controversial, bearing on the comparative physiology and ecology of locomotion^{1,2}. Applied to flying birds, aerodynamic theory predicts that mechanical power should vary as a function of forward velocity in a U-shaped curve. The only empirical test of this theory, using the black-billed magpie (*Pica pica*), suggests that the mechanical power curve is relatively flat over intermediate velocities³. Here, by integrating *in vivo* measurements of pectoralis force and length change with quasi-steady aerodynamic models developed using data on wing and body movement, we present mechanical power curves for cockatiels (*Nymphicus hollandicus*) and ringed turtle-doves (*Streptopelia risoria*). In contrast to the curve reported for magpies³, the power curve for cockatiels is acutely concave, whereas that for doves is intermediate in shape and shows higher mass-specific power output at most speeds. We also find that wing-beat frequency and mechanical power output do not

necessarily share minima in flying birds. Thus, aspects of morphology, wing kinematics and overall style of flight can greatly affect the magnitude and shape of a species' power curve.

We integrated several techniques to study *in vivo* mechanical power output from the pectoralis, the primary flight muscle in birds, as a function of forward flight velocity. We flew cockatiels and doves in a variable-speed wind tunnel and measured wing and body movements in three dimensions simultaneously with direct measures of force development and length change of the pectoralis. The three-dimensional (3D) kinematic data were incorporated into a model of quasi-steady aerodynamic power output, which we developed using existing theory⁴⁻⁶. Pectoralis tension and length change were measured using bone-strain recordings and sonomicrometry, respectively (ref. 7 and Fig. 1a). Our aerodynamic model, coupled with observed changes in muscle length, was used to calibrate measurements of bone strain into units of muscle force at intermediate speeds (7 and 9 m s⁻¹). These force calibrations were then applied to bone-strain data at all flight speeds.

To estimate work output per wing beat, we integrated tension-length curves, or work loops, for each wing beat (Fig. 1b). Average power output from the pectoralis muscles was then calculated as total work during an interval of flight divided by flight duration. Pectoralis power output varied significantly with flight speed in both species (repeated-measures analysis of variance, $P \leq 0.002$; Fig. 2a, b). In the cockatiels, minimum power output was 1.3 W (73.9 W kg⁻¹ muscle mass) at 5 m s⁻¹, and maximum power output was 3.7 W (231.2 W kg⁻¹) at 14 m s⁻¹, the fastest speed at which the cockatiels would fly. In the doves, minimum power (4.3 W;

123.4 W kg⁻¹) was observed at 7 m s⁻¹, and maximum power output (7.5 W; 234.5 W kg⁻¹) was observed at this species' maximum flight speed of 17 m s⁻¹. Mass-specific power output in the dove was higher than that for the cockatiel at all speeds except 14 m s⁻¹. This is probably due to the higher weight per area swept by the wings (disc loading) of the dove (8.1 N m⁻²) as compared with the cockatiel (4.2 N m⁻²) and therefore to the proportionally greater induced power requirements of the dove.

Our results do not support a current hypothesis that minimum wing-beat frequency must coincide with a bird's minimum power speed⁸. For cockatiels and doves, relative work varied closely with relative power, and both variables showed shared minima (Fig. 3b, c). By contrast, wing-beat frequency in cockatiels reached a well-defined minimum at 9 m s⁻¹, a speed 80% higher than their minimum power speed, whereas wing-beat frequency in doves showed a broad minimum from 7 to 13 m s⁻¹.

Power curves in doves and cockatiels differed substantially from the curve measured in magpies, the only other species for which similar muscle measurements have been made over a range of flight

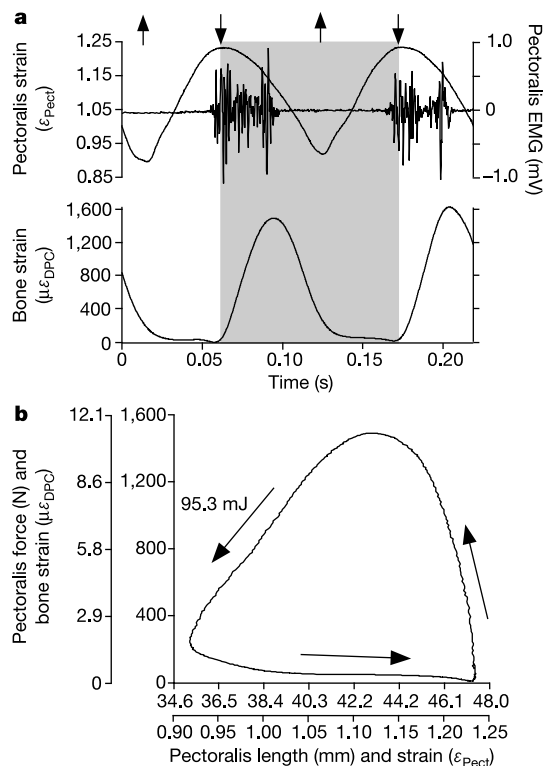


Figure 1 *In vivo* recordings of pectoralis strain, force and work per wing-beat cycle. **a**, Changes in muscle length, electromyographic (EMG) activity and bone strain recorded from a cockatiel flying at 7 m s⁻¹. Shaded region indicates a single wing-beat cycle selected for plotting as a tension-length curve (work loop in **b**). Arrows indicate upstroke and downstroke. **b**, Work loop plotted from a single wing beat (shaded region in **a**). We calibrated bone and pectoralis strain into pectoralis force and length using our aerodynamic model and the bird's morphology. The integral of the work loop (95.3 mJ) was divided by time (112.2 ms) to yield power for a single pectoralis (0.85 W), and this value was doubled to yield total pectoralis power (1.7 W or 106 W kg⁻¹ pectoralis mass) for the wing beat.

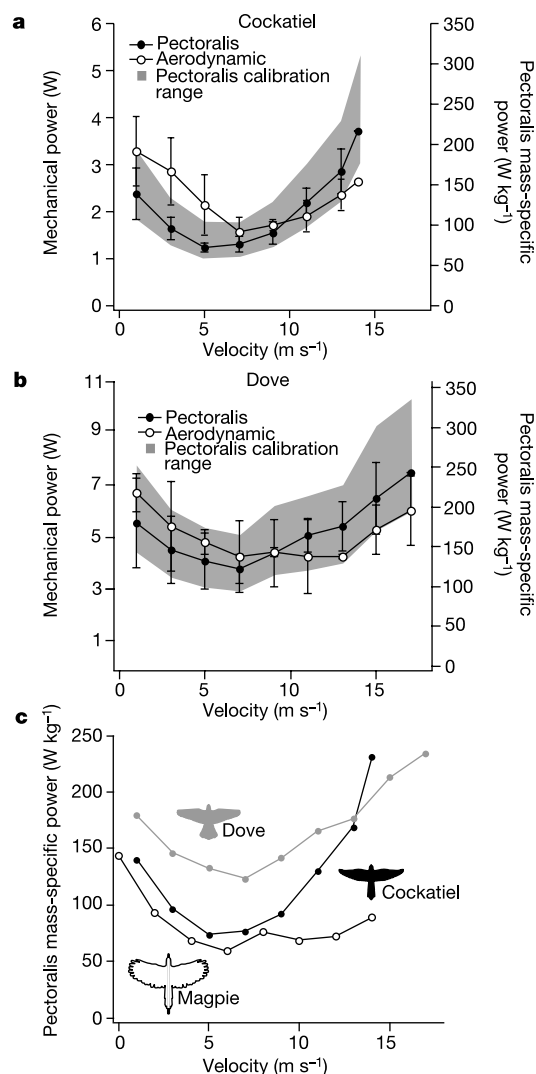


Figure 2 Pectoralis power as a function of flight velocity. **a**, **b**, Mean $\pm$ s.d. pectoralis and aerodynamic power as a function of flight velocity in five cockatiels (**a**) and three doves (**b**). In both cases, the shaded region indicates the range of pectoralis power obtained when 'low-power' versus 'high-power' aerodynamic coefficients are input to the aerodynamic model (Methods). **c**, Comparative mass-specific pectoralis power as a function of flight velocity in cockatiels, doves and magpies³. Bird silhouettes are shown to scale, digitized from video.

speeds³. Differences in calibration method for strain-gauge measurements cannot account for all of the interspecific differences because relative power curves varied between the species (Fig. 3a), and relative power and work (Fig. 3b, c) were independent of force calibration.

As compared with the mechanical power curve for the magpie, which has a broadly flat shape between 4 and 12 m s⁻¹, progressively greater upward concavity was apparent for the curves shown by doves and cockatiels, respectively (Figs 2c and 3a). Maximum mass-specific pectoralis power output in cockatiels and doves was 63% higher than the maximum of 143 W kg⁻¹ previously reported for magpies³. Unlike the cockatiels and doves, magpies are reported to generate maximal pectoralis power during hovering³. We attribute these differences in the shape and magnitude of the power curves to differences in wing and tail morphology.

Magpies have broad, rounded wings (aspect ratio 5) and a long tail. This probably increases profile drag, diminishes thrust/drag ratio² and prevents them from flying faster than 14 m s⁻¹, even though they seem to have sufficient pectoralis power to fly at faster speeds. By contrast, cockatiels and doves have relatively pointed wings of higher aspect ratio (7.0 and 5.7) and proportionally smaller tails. These characteristics offer proportionally lower induced velocities and profile drag and, thus, more favourable thrust/drag ratios, permitting them to fly at speeds that require pectoralis power

output in excess of that required for hovering. Unlike in magpies, maximum speed in cockatiels and doves is probably constrained by sustainable power available from the pectoralis muscle.

Differences in power curves between magpies and the species in the present study may also be explained by flight style. Magpies use a unique intermittent flight style in which they regularly fluctuate their wing-beat gait, flight velocity and altitude⁹. This variation in wing motion may alter power requirements at intermediate and fast speeds so that the right side of the power curve remains relatively flat for this species.

Maximum mass-specific pectoralis power output in the cockatiel and dove was 60% or less than values (360–460 W kg⁻¹) that have been estimated for galliform species (pheasant-like birds) engaged in short-duration take-off and climbing flight⁶. In those studies, flight style and speed range differed, but similar aerodynamic modelling and sonomicrometry methods were used. Thus, we attribute most of the interspecific variation between these studies to differences in non-sustainable versus sustainable flight performance and flight morphology. Among all species of birds, galliforms seem to be designed for maximum burst-flight performance: they have short, broad wings, high wing-beat frequency and pectoralis muscles with fast myosin ATPase and high glycolytic capacity^{10–12}. They use flight almost only as an escape mechanism, rapidly fatiguing after take-off and returning to the ground to run or hide, whereas cockatiels and doves may sustain flight for hours before fatigue.

Data from the cockatiels and doves provide insight into the validity of certain assumptions in quasi-steady aerodynamics as applied to bird flight. Because we used aerodynamic estimates of power output at intermediate flight speeds to calibrate bone strain into muscle force, by necessity the curves for aerodynamic power and pectoralis power intersect in the vicinity of 7–9 m s⁻¹ for both species (Fig. 2a, b). But pectoralis power was not constrained to match aerodynamic power at slower or faster speeds. It is therefore significant that our measurements of aerodynamic power overestimate pectoralis power at slow speeds and underestimate pectoralis power at fast speeds in both species. Several explanations could account for these discrepancies. Wing muscles other than the pectoralis may contribute significantly to work and power during slow flight¹³, and we did not measure power output in these muscles. In addition, our model accounts only for the wings; other lift surfaces, especially the tail, may be particularly important for reducing power requirements at slow speeds^{14,15}. By contrast, at faster speeds our model seems to underestimate drag. Drag coefficients on bird wings and bodies are subject to much uncertainty⁸. Without more definitive estimates for these parameters and induced velocity, we are unable to determine which components of total drag were underestimated.

In summary, our data show that differences in morphology, wing kinematics and overall style of flight have large effects on the magnitude and shape of a species' power curve. □

Methods

Birds, kinematic analysis and aerodynamic modelling

Five cockatiels (body mass 78.5 g, wing span 48.3 cm) and three doves (139.8 g, 46.8 cm) were trained to fly individually in a variable-speed wind tunnel at the Concord Field Station, Harvard University (41-m altitude, average air density (ρ) 1.22 kg m⁻³). We followed Institutional Animal Care and Use guidelines in all housing and experiment protocols. The working-section diameter of the tunnel was 1.2 × 1.2 m (ref. 15). We report equivalent air speed V_e rather than true air speed¹⁶:

$$V_e = \sqrt{2q/\rho_0} \quad (1)$$

where q is dynamic pressure and ρ_0 is the air density at sea level (1.225 kg m⁻³). The birds were trained to sustain flight indefinitely at intermediate speeds (5–11 m s⁻¹) and for at least 20 s at other speeds. Kinematic data were obtained using 2–4 synchronized, 250-Hz digital video cameras ($n = 214$ wing beats for cockatiels, $n = 98$ for doves). We merged two-dimensional coordinates from each camera plane into a single 3D coordinate space using the direct linear transformation (DLT) coefficients derived from a 40-point calibration frame¹⁵. Variables obtained from the kinematic analyses were used in constructing quasi-steady aerodynamic models incorporating actuator-disc theory and

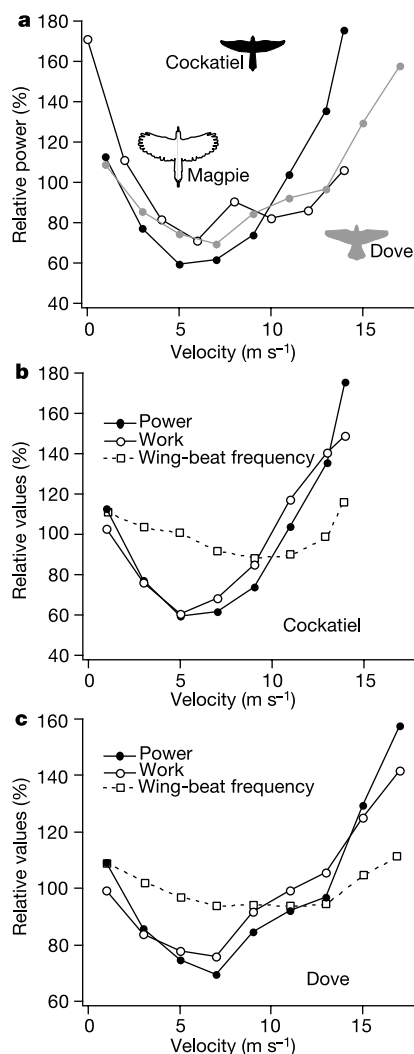


Figure 3 Pectoralis power, work and wing-beat frequency as a function of flight velocity, with values expressed as a percentage of the observed within-species mean for each variable. **a**, Mean relative pectoralis power in cockatiels, doves and magpies⁷. **b, c**, Mean relative pectoralis power, work and wing-beat frequency in cockatiels (**b**) and doves (**c**).

strip analysis⁴⁻⁶. Relative to free-stream velocity, we measured body angle, wing span, wing elevation, stroke-plane angle of the wing and angular velocity of the wing. Total aerodynamic power (P_{aero}) was calculated as the sum of induced P_{ind} , profile P_{pro} and parasite P_{par} power, and rate change in potential energy. Inertial power, added mass and rate change in kinetic energy were ignored. For the aerodynamic variables contributing to P_{aero} :

$$P_{\text{ind}} = M_b k w (g + A_v) \quad (2)$$

where M_b is body mass, k is the induced velocity correction factor (assumed to be 1.2), w is induced velocity, g is gravitational acceleration, and A_v is vertical acceleration. In our calculation of w , we defined disc area as the area swept by the wings, not a circle⁶.

$$P_{\text{pro}} = 2 \sum_{i=1}^{25} \left(\frac{1}{2} \rho V_{R,i} S_i C_{D,\text{pro}} \right) \quad (3)$$

where V_R is the resultant velocity of a given wing strip, S is the surface area of a wing strip, and $C_{D,\text{pro}}$ is the profile drag coefficient (assumed to be 0.02)^{4,5}. We assumed no change in average wing chord throughout the wing-beat cycle.

$$P_{\text{par}} = \frac{1}{2} \rho S_b C_{D,\text{par}} V_e^2 \quad (4)$$

where S_b is the projected equivalent flat-plate area of body, calculated incorporating body angle relative to free-stream flow⁴, and $C_{D,\text{par}}$ is the parasite drag coefficient (assumed to be 0.13)^{2,6}.

Bone strain and muscle length

Using surgically implanted strain gauges and sonomicrometry crystals, we obtained *in vivo* measurements of bone strain (ϵ_{DPC}) and pectoralis strain (ϵ_{pect} ; Fig. 1a). Strains were defined with resting (initial) length as that during perching with the bird motionless. A single-element, 1-mm metal-foil strain gauge was placed perpendicular to the long axis of the humerus on the dorsal surface of the delto-pectoral crest (DPC), the insertion for the pectoralis. We assumed that ϵ_{DPC} is proportional to muscle tension. Two 1-mm sonomicrometry crystals (resolution $\pm 15 \mu\text{m}$) were placed 2.5-mm deep, 15-mm apart into the longest fascicles of the mid-belly, the pars sternobrachialis of the pectoralis. We used indwelling electromyography electrodes to measure the timing of pectoralis activity (Fig. 1a).

Calculating pectoralis work and power

Within each wing beat sampled ($n = 979$ wing beats for cockatiels, $n = 1,067$ for doves), we integrated the curve of ϵ_{DPC} plotted as a function of ϵ_{pect} . The area of these dimensionless loops was converted into work (J) per wing beat by multiplying ϵ_{pect} by L_{rest} the resting fascicle length (m) of the pectoralis in the region implanted with sonomicrometry crystals, and by multiplying ϵ_{DPC} by F , a force calibration factor (Fig. 1b). We avoided exclusive reliance on our earlier "pull-calibration" technique⁷ because we were unable to obtain consistent calibrations in individuals of both species: pull calibrations yielded an average coefficient of variance for F of $29.0 \pm 14.7\%$ in individual birds. Owing to the superficial application of tensile force, this approach also probably engenders larger bending moments than *in vivo* forces transmitted by the muscle at its attachment site, leading to an underestimate of F . Because of these concerns, F was calculated so that mechanical power in the paired pectoralis was set to equal P_{tot} for the subset of wing beats for which kinematics were analysed at 7 and 9 m s⁻¹ ($n = 63$ for cockatiels, $n = 29$ for doves). These speeds were chosen because flight at slower speeds is more likely to violate our assumption of quasi-steady flow¹³, and P_{aero} at faster speeds is more sensitive to assumed values of $C_{D,\text{pro}}$ and $C_{D,\text{par}}$. For each bird, pectoralis power, P_{pect} , was calculated as total work output at a given speed divided by total time in flight at that speed. This approach yielded more consistent calibrations ($\text{cv} = 10.2 \pm 4.2\%$) for F in individual birds.

To evaluate the sensitivity of our aerodynamic-calibration approach, we varied the values of k , $C_{D,\text{pro}}$ and $C_{D,\text{par}}$ for a low- versus a high-power case. In the low-power case, k was decreased from 1.2 to 1.0, and $C_{D,\text{par}}$ and $C_{D,\text{pro}}$ were decreased by 50% to 0.065 and 0.01, respectively; for the high-power case, k was assumed to be 1.4, and $C_{D,\text{par}}$ and $C_{D,\text{pro}}$ were increased by 50% to 0.195 and 0.03. This allowed us to bracket the likely range of pectoralis power that might be observed at any particular flight speed (Fig. 2a, b). We also computed relative power, work and wing-beat frequency, all independent of F , by dividing the observed values in physical units by their respective within-species means and multiplying the dimensionless results by 100 to obtain percentages (Fig. 3).

Received 29 July; accepted 29 October 2002; doi:10.1038/nature01284.

1. Ellington, C. P. Limitations on animal flight performance. *J. Exp. Biol.* **160**, 71–91 (1990).
2. Rayner, J. M. V. Estimating power curves of flying vertebrates. *J. Exp. Biol.* **202**, 3449–3461 (1999).
3. Dial, K. P., Biewener, A. A., Tobalske, B. W. & Warrick, D. R. Mechanical power output of bird flight. *Nature* **390**, 67–70 (1997).
4. Rayner, J. M. V. A vortex theory of animal flight. Part 2. The forward flight of birds. *J. Fluid Mech.* **91**, 731–763 (1979).
5. Norberg, U. M. *Vertebrate Flight: Mechanics, Physiology, Morphology, Ecology and Evolution* (Springer, Berlin, 1990).
6. Askew, G. N., Marsh, R. L. & Ellington, C. P. The mechanical power output of the flight muscles of blue-breasted quail (*Coturnix coturnix*) during take-off. *J. Exp. Biol.* **204**, 3601–3619 (2001).
7. Biewener, A. A., Corning, W. R. & Tobalske, B. W. *In vivo* muscle force-length behavior during level flight in pigeons (*Columba livia*). *J. Exp. Biol.* **201**, 3293–3307 (1998).
8. Pennycuik, C. J., Klaassen, M., Kvist, A. & Lindström, Å. Wingbeat frequency and the body drag anomaly: wind-tunnel observations on a thrush nightingale (*Luscinia luscinia*) and a teal (*Anas crecca*). *J. Exp. Biol.* **199**, 2757–2765 (1996).
9. Tobalske, B. W. Biomechanics and physiology of gait selection in flying birds. *Physiol. Biochem. Zool.* **73**, 736–750 (2000).

10. Tobalske, B. W. & Dial, K. P. Effects of body size on take-off performance in the Phasianidae (Aves). *J. Exp. Biol.* **203**, 3319–3332 (2000).
11. Rosser, B. W. C. & George, J. C. The avian pectoralis: histochemical characterization and distribution of muscle fiber types. *Can. J. Zool.* **64**, 1174–1185 (1986).
12. Rosser, B. W. C., Wick, M., Waldbillig, D. M. & Bandman, E. Heterogeneity of myosin heavy-chain expression in fast-twitch fiber types of mature avian pectoralis muscle. *Biochem. Cell Biol.* **74**, 715–728 (1996).
13. Dial, K. P. Activity patterns of the wing muscles of the pigeon (*Columba livia*). *J. Exp. Zool.* **262**, 357–373 (1992).
14. Thomas, A. L. R. On the aerodynamics of birds' tails. *Phil. Trans. R. Soc. Lond. B* **340**, 361–380 (1993).
15. Hedrick, T. L., Tobalske, B. W. & Biewener, A. A. Estimates of circulation and gait change based on a three-dimensional kinematic analysis of flight in cockatiels (*Nymphicus hollandicus*) and ringed turtle-doves (*Streptopelia risoria*). *J. Exp. Biol.* **205**, 1389–1409 (2002).
16. Pennycuik, C. J., Alterstam, T. & Hedenström, A. A new low-turbulence wind tunnel for bird flight experiments at Lund University, Sweden. *J. Exp. Biol.* **200**, 1441–1449 (1997).

Acknowledgements We thank G. Lauder and F. Jenkins Jr for comments on this work, and the Concord Field Station staff for animal care and help with experiments. This work was supported by grants from the National Science Foundation to A.A.B. and K.P.D., and from Murdock to B.W.T.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.W.T. (e-mail: tobalske@up.edu).

Neuronal synchrony does not correlate with motion coherence in cortical area MT

Alexander Thiele*† & Gene Stoner*

* Salk Institute for Biological Studies, La Jolla, California 92037, USA

Natural visual scenes are cluttered with multiple objects whose individual features must somehow be selectively linked (or 'bound') if perception is to coincide with reality. Recent neurophysiological evidence^{1,2} supports a 'binding-by-synchrony' hypothesis³: neurons excited by features of the same object fire synchronously, while neurons excited by features of different objects do not. Moving plaid patterns offer a straightforward means to test this idea. By appropriate manipulations of apparent transparency, the component gratings of a plaid pattern can be seen as parts of a single coherently moving surface or as two non-coherently moving surfaces. We examined directional tuning and synchrony of area-MT neurons in awake, fixating primates in response to perceptually coherent and non-coherent plaid patterns. Here we show that directional tuning correlated highly with perceptual coherence, which is consistent with an earlier study⁴. Although we found stimulus-dependent synchrony, coherent plaids elicited significantly less synchrony than did non-coherent plaids. Our data therefore do not support the binding-by-synchrony hypothesis as applied to this class of motion stimuli in area MT.

To confirm that monkeys' perception is influenced by transparency manipulations⁵⁻⁸, we trained one monkey to distinguish coherent from non-coherent motion. After training, perceptual reports were obtained for nine probe plaids. For eight of these plaids, we manipulated transparency by holding the intensities of the two gratings (thin bar phase) constant and identical, and varying the intensity of the grating intersections⁵⁻⁸. The ninth probe plaid had one bright and one dark grating. We termed this

† Present address: Henry Wellcome Building for Neuroecology, University of Newcastle upon Tyne, NE2 4HH, UK.

plaid 'depth-ordered non-coherent' because the bright bars appeared to transparently occlude the dark bars owing to the intersection luminance.

Monkey coherency judgements (Fig. 1a) matched those of humans⁵. The plaid intersection luminance yielding the highest percentage (97%) of non-coherent motion reports was consistent with transparency. Conversely, the luminance yielding the smallest percentage (31%) of non-coherent motion reports was inconsistent with transparency. For depth-ordered non-coherent plaids, 99% of trials were reported as non-coherent (see Supplementary section 1).

On the basis of our psychophysical results, we selected three plaids to test the binding-by-synchrony (BBS) hypothesis. Coherent plaids (Fig. 1b, right) had an intersection luminance that yielded mostly coherent reports. Our non-coherent plaid (Fig. 1b, left), conversely, had an intersection luminance that consistently resulted in non-coherent reports. We also used depth-ordered non-coherent plaids, which humans and our monkey subject usually saw as non-coherent (Fig. 1b, centre). Thus, we used one perceptually coherent plaid and two types of non-coherent plaids in our neurophysiological studies.

We recorded neuronal activity from pairs of directionally selective neurons in macaque area MT from three fixating monkeys (one of which had previously engaged in the behavioural experiments). Applied to motion coherency, the BBS hypothesis implies that two neurons, one preferring the direction of motion of one grating, the other preferring the direction of motion of the other grating, should synchronize for perceptually coherent but not for non-coherent motion⁹.

The majority of our neuronal pairs did not conform to this prediction. Many neuronal pairs exhibited synchrony that was dependent on the direction of motion of individual gratings (Fig. 2a, left panels) and of plaids (Fig. 2a, right panels). Maximal synchrony was often associated with directions of plaid motion other than those central to the BBS hypothesis (Fig. 2a). Moreover, synchrony seen for the directions relevant to the BBS hypothesis was seldom largest for coherent plaids (for example, see Fig. 2).

Our measure of synchrony was the neuronal correlation coefficient (NCC)¹⁰. The larger the NCC, the greater the proportion of spikes with nearly synchronous timing. Figure 3a shows NCCs cross-plotted for the three pairwise plaid comparisons (143 cell pairs; time period analysed: 200–2,000 ms after stimulus onset). Contrary to the BBS prediction, synchrony elicited by our coherent plaids was statistically indistinguishable from that of non-coherent plaids. Surprisingly, synchrony elicited by depth-ordered non-coherent plaids was significantly greater than that for either non-coherent ($P = 0.012$, repeated measurement (RM) analysis of variance (ANOVA), Tukey's test) or coherent plaids ($P < 0.001$).

To determine whether firing rate was related to coherency and/or synchrony, we calculated average firing rates on the basis of the same neuronal responses used to calculate NCCs and made the same pairwise comparisons. Firing rates elicited by depth-ordered non-coherent and coherent plaids (for which the pattern direction was around 67.5° away from the preferred direction of each neuron) were significantly greater than those elicited by non-coherent plaids (Fig. 3b). Thus, firing rate discriminated different plaids than did synchrony, and neither measure distinguished coherent from both types of non-coherent plaids. Although both neuronal synchrony and firing rates (obtained from plaid motion that is central for testing the BBS) failed to correlate consistently with perceptual coherence, plaid pattern directional tuning curves (based on all eight directions of plaid motion) correlated well. A single measure, the component pattern index (CPI) captures the degree to which these tuning curves reflected responses to the two individual gratings rather than to the pattern as a whole^{4,11,12}. Positive CPI values imply the encoding of two motions, and negative CPI values suggest a single motion representation. Typically, the larger the CPIs the more bi-lobed is the tuning curve. Figure 3c shows CPIs

calculated for the 225 individual neurons that were members of the neuronal pairs used in our synchrony analysis. Both types of non-coherent plaid elicited CPIs that were significantly greater than for the coherent plaid, whereas no significant difference existed between non-coherent and depth-ordered non-coherent plaids. These results confirm that the directional tuning of area MT neurons is correlated with perceptual coherence⁴.

Recently, a related study came to differing conclusions¹³. They report that cortical neurons in visual areas A18 and PMLS of anaesthetized cats synchronize their responses as a function of perceptual motion coherence but show no coherency-related changes in directional tuning. Although interspecies differences could account for the apparent discrepancy, differences in state of consciousness could also have a role. In monkeys, anaesthesia has been reported to disrupt the ability of directionally selective neurons to integrate motion¹⁴. Furthermore, we examined only synchrony between directionally selective neurons with directional preferences near to the direction of motion of the plaid's two gratings, whereas ref. 13 did not separately present data from neuronal pairs satisfying that restriction. Because of these profound differences, the two studies cannot be compared directly. To determine whether synchrony related to binding had escaped our analyses, we investigated several variations of the BBS hypothesis. It is possible, for example, that synchrony is expressed differently in different classes of directionally selective neurons. On the basis of responses to coherent plaids, directionally selective neurons have been classified into two types¹¹: 'component' neurons respond to the motions of the individual gratings, and 'pattern' neurons respond to the motion of the plaid as a whole. Pattern neurons are thought to

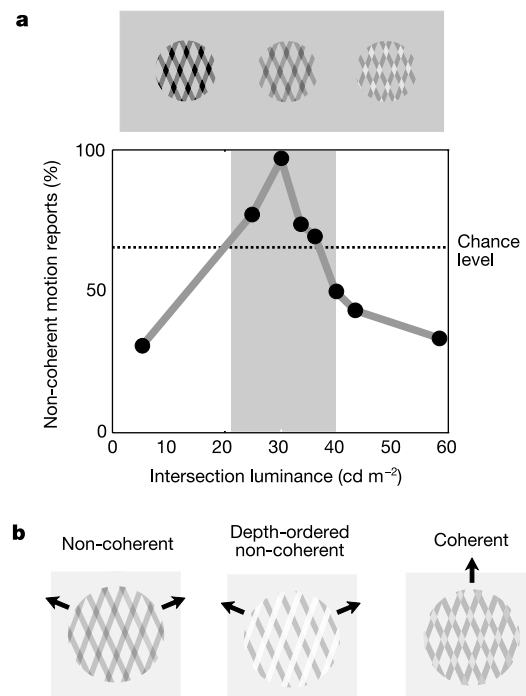


Figure 1 Stimulus and monkey psychophysics. **a**, Monkey perceptual reports as a function of luminance configuration (35 trials per stimulus). The intersection luminance of the probe stimuli (x axis) was varied. Luminance values lying within the central region (grey panel) correspond to stimuli that, when stationary, human subjects usually interpret as one transparent surface overlapping another. When applied to moving plaid patterns, human subjects see maximal non-coherency within this transparency zone²¹. Monkey behavioural data matched that found for humans^{5,21}. Stimulus icons above the graph approximate the appearance of plaids corresponding to the different intersection luminance values. **b**, Three plaids used in neurophysiological experiments. Non-coherent and depth-ordered, non-coherent plaids are seen as two gratings moving independently. Coherent plaids are seen as a single moving object. Arrows indicate perceived motions.

achieve their response properties through component neuron input. Castelo-Branco *et al.*¹³ recorded almost exclusively from component neurons, whereas primate area MT has both component and pattern neurons^{4,11,12}. Rate modulation and synchronization may coexist and complement each other^{9,13} such that synchrony among input neurons determines, in part, the firing rate of recipient higher-order neurons.

On the basis of these assumptions, a scenario consistent with both BBS and the two-stage (component–pattern neuron) model is that component neurons show coherency-related synchrony, whereas pattern neurons show only changes in directional tuning (that is, rate changes for specific directions). If temporal synchrony occurs among component but not pattern neurons, then the pattern neurons in our data set might have masked the synchrony effects of the component neurons. To investigate this possibility, we subdivided neurons into component and pattern types. We analysed separately synchrony between component–component, component–pattern, and pattern–pattern neuronal pairs. For the com-

ponent–component group, depth-ordered non-coherent plaids elicited greater average synchrony than either non-coherent or coherent plaids ($n = 45$, $P < 0.05$, RM-ANOVA, Tukey's test). Although the trend was similar for pattern–pattern and component–pattern neuronal pairs it was not significant, possibly owing to the small sample sizes ($n = 17$, and $n = 28$).

We also examined coherency-related changes in CPI values for pattern and component populations independently. Consistent with previous reports⁴, both neuronal types become significantly ($P < 0.001$) more component-like (their CPI values become larger) when stimulated with perceptually non-coherent plaids compared to coherent plaids. Thus our data do not support a hierarchical binding model in which component neuron synchrony induces changes in the directional tuning of pattern neurons. On the contrary, directional tuning correlates of perceptual coherency are found within both populations of area MT neurons and synchrony consistent with BBS is found in neither.

Another possibility is that BBS occurs in the ON response

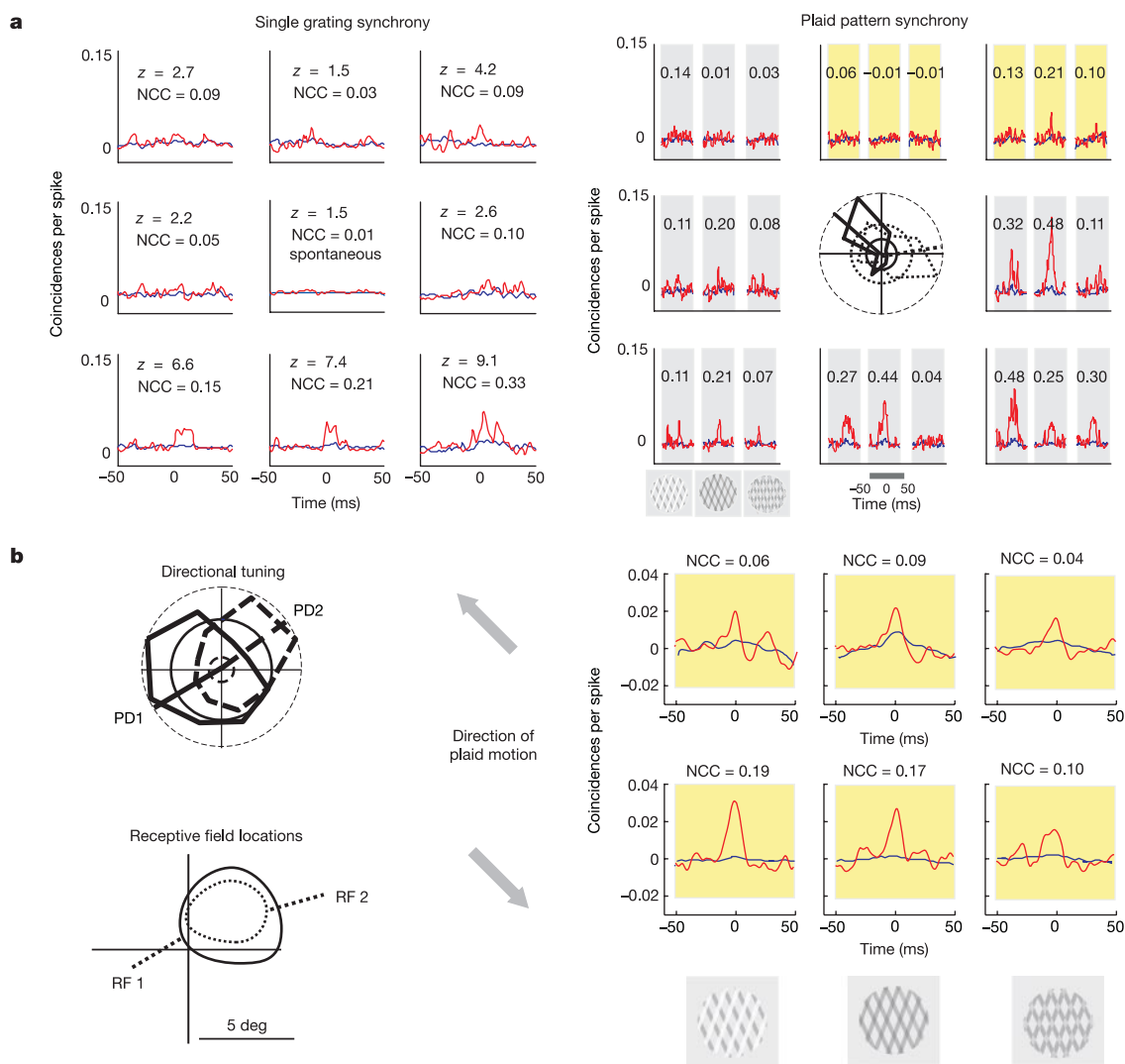


Figure 2 Data from two neuronal pairs. For both pairs, two directions of plaid motion were suitable for testing the binding-by-synchrony (BBS) hypothesis (relevant correlograms are highlighted in yellow). **a**, Responses of a neuronal pair to gratings (left) and plaids (right). Positioning of cross-correlograms (red) and peristimulus time histogram (PSTH) predictors (blue) reflects direction of stimulus motion (left correlogram indicates leftward motion, and so on). Centre inset of right panel shows grating directional tuning curves. Three correlograms in each of the eight subplots of the right panel are for depth-ordered non-coherent, non-coherent, and coherent plaids (from left to right). Neuronal correlation

coefficients (NCCs) are shown above each correlogram. Left panel also shows peak z -scores (z). Synchrony was either greatest for non-coherent plaids or absent altogether. **b**, The second pair showed the least synchrony for coherent plaids. Grating directional tuning curves and receptive fields are shown. For all correlograms, the x axis indicates time interval between spikes from two neuronal recordings and the y axis the number of coincidences. All firing rates are normalized to individual maxima. For tuning curves, dashed and solid lines distinguish individual neurons and vectors indicate the preferred direction of motion (inner dashed and solid circles show spontaneous activity).

transient but that binding is later expressed in firing rates (see ref. 15 for a related finding). To examine this possibility we analysed correlated activity over time periods that included the ON response transient: 30–200 and 30–2,000 ms relative to stimulus onset. The time period of 30–200 ms yielded results qualitatively similar to our initial analysis with depth-ordered non-coherent stimuli eliciting the most synchrony (see Supplementary section 2). However, for this short time period, that difference was only significant when comparing depth-ordered non-coherent and non-coherent stimuli. The time period of 30–2,000 ms yielded results nearly identical to those found for the period after the ON response transient (200–2,000 ms). Thus BBS is found in neither the ON response transient nor the sustained stimulus-response period.

A further possibility is that synchrony serves only to bind features encoded by different hypercolumns⁹. Because the receptive fields of many of our neurons were nearly spatially coincident, both members of some neuronal pairs might be part of the same direction-of-

motion hypercolumn. We divided our neuronal pairs into those with less than 50% receptive field overlap and those with greater than 50% overlap (mean receptive field overlap: $45.4 \pm 48.9\%$, range: 100%; mean receptive field separation: $1.91 \pm 1.71^\circ$, range: 7.61° ; for details see Supplementary section 3). For both sets of neuronal pairs, the depth-ordered stimulus always induced the largest amount of synchrony. However, synchrony differences were slightly greater for cell pairs with non-overlapping receptive fields ($n = 57$, depth-ordered versus coherent: $P = 0.003$; depth-ordered versus non-coherent: $P = 0.037$; coherent versus non-coherent: $P = 0.697$) than for pairs with overlapping receptive fields ($n = 86$, depth-ordered versus coherent: $P = 0.127$; depth-ordered versus non-coherent: $P = 0.021$; coherent versus non-coherent: $P = 0.243$; RM-ANOVA, Tukey's test).

Finally, we examined the contribution of eye movements to neuronal synchrony. Small eye movements (mini-saccades and smooth tracking) unavoidably accompany conscious viewing even during fixation. Tracking eye movements tend to increase perceptual coherence (G.S., unpublished observations) and hence dilute the perceptual differences between stimuli that are intended to be coherent and non-coherent. In addition, eye movements may synchronously activate (or suppress) MT neurons causing stimulus-related changes in neuronal correlation. For these reasons, we analysed data from time periods in which eye movements were absent or minimal. Surprisingly, this analysis revealed that both depth-ordered non-coherent ($P = 0.004$) and non-coherent plaids ($P = 0.01$) elicited significantly more synchrony than did coherent plaids (RM-ANOVA, Tukey's test). For these eye-movement-free periods, no statistical difference in synchrony was found between non-coherent and depth-ordered non-coherent plaids. Thus, during periods of viewing best suited to testing the BBS hypothesis (in which perceptual differences between plaid stimuli are likely to be maximal), we find a result opposite to that predicted by the BBS hypothesis: correlated firing decreases when moving features are perceptually bound into a coherent whole.

Thus we have confirmed that perceptual motion coherency is accompanied by changes in directional tuning within area MT but have found no support for the hypothesis that binding is accompanied by increased neuronal synchrony. Stimulus-induced synchrony was in fact greatest, not for coherent plaids, but for non-coherent plaids. The functional significance, if any, of this synchrony awaits further research. □

Methods

General

We recorded from 339 directionally selective cell pairs from area MT in three fixating rhesus monkeys (*Macaca mulatta*). Of these pairs, 143 had directional preferences suitable to testing the BBS hypothesis (sampled from 36, 24 and 23 sites in the three monkeys, respectively). We used a linear four-electrode device (electrode positioning system, Alpha Omega) attached to a custom-built head stage (electrode spacing: $350 \mu\text{m}$ – $800 \mu\text{m}$ – $350 \mu\text{m}$). Single units (96 pairs were single units) were isolated using spike-sorting techniques (Alpha Omega). Multi-unit activity (47 pairs) criteria were standard¹⁶ as were the neurophysiological and animal training methods^{15,17}. Experimental protocols conformed to US Department of Agriculture (USDA) regulations and National Institutes of Health (NIH) guidelines for the humane care and use of laboratory animals.

Behavioural task and visual stimuli

Animals fixated a 0.4° diameter target (within ± 1.0 – 1.5°) during 700-ms pre-stimulus and 2,000-ms stimulus presentation period. Stimuli were generated on a cathode ray tube (refresh rate: 60 Hz for a few recordings, 100 Hz otherwise). Plaid stimuli consisted of two superimposed gratings moving at an angle of 135° relative to each other within a circular aperture of 10° . The stimulus was either centred on the fixation spot (97% of the cases) or within 3° of the fixation spot (for extended Methods see Supplementary section 4). In electrophysiological experiments, we used coherent, non-coherent, and depth-ordered non-coherent plaids (for additional details, see Supplementary section 5). Neuronal responses and synchrony were assessed using eight different directions of plaid motion (0 – 315° , in steps of 45°) and eight directions of single grating motion (corresponding to those in which the plaid components moved; that is 22.5 – 337.5° , in steps of 45°). Grating duty cycle (thin bar phase grating period) was 0.24. Grating speed was 4° s^{-1} and the plaid pattern speed was 10° s^{-1} . Neural synchrony was analysed if a cell pair's preferred direction differed by $\geq 90^\circ$ and plaids moved intermediate to the preferred directions of the two cells

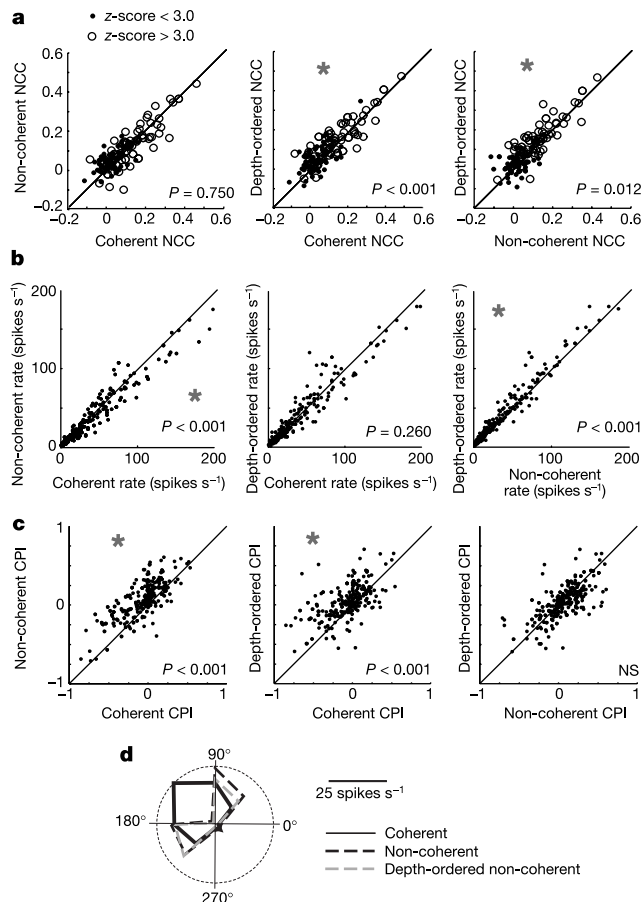


Figure 3 Synchrony magnitude firing rate and component-pattern indices cross-plotted for each pair-wise stimulus comparison. **a**, Neuronal correlation coefficients (NCCs) were significantly larger for depth-ordered non-coherent than for coherent and non-coherent stimuli. No significant synchrony difference was found between non-coherent and coherent stimuli. Open symbols: z-score synchrony measure exceeded 3.0 for at least one stimulus. Filled symbols: z-score < 3.0 for all stimuli. **b**, Firing rates of individual cells based on the same neuronal responses used to compute synchrony. Rates for depth-ordered non-coherent and coherent plaids were both significantly higher than for non-coherent plaid. **c**, Component pattern index (CPI) values for both non-coherent and depth-ordered non-coherent were significantly greater than for coherent stimuli. **d**, Directional tuning of MT neuron to three plaids. Tuning curves are more component-like for non-coherent and depth-ordered non-coherent plaids than for coherent plaids. For all stimulus comparisons (**a**–**c**), asterisks indicate significant differences and are positioned on the side of the diagonal with greater magnitude (reflected in the offset of data points).

(average directional preference difference: $138.3 \pm 26.2^\circ$, range: 89.1°). For cell pairs in which two directions of plaid motion fulfilled these selection criteria, neuronal correlation coefficients (NCCs) were averaged.

Perceptual reports

One monkey reported the direction of moving plaid patterns (six-alternative forced-choice task). During training, the animal viewed plaid patterns with a random-dot texture 'pasted' on the thin bar phase of the gratings (for demonstrations and details see Supplementary section 6). The texture on the two gratings either moved in the same direction (unambiguous coherent motion) or in different directions (unambiguous non-coherent motion). The monkey had one valid choice target for coherent motion and two valid choice targets for non-coherent motion (80% rewards on correct choices). Once performance had reached around 85% correct, we introduced ambiguous probe plaid stimuli on 20% of the trials. On these trials, the monkey was rewarded for reports consistent with either coherent or non-coherent motion.

Data analysis

Baseline firing rates and neuronal synchrony were calculated from 200 ms after fixation onset until stimulus onset (700 ms after fixation onset). Stimulus related activity and synchrony were calculated for three periods: (1) from 200 ms after stimulus onset until 2,000 ms thereafter, avoiding the ON response transient; (2) from the earliest response (30 ms after stimulus onset) until 2,000 ms after stimulus onset; and (3) for the ON response transient only (30–200 ms after stimulus onset). Preferred directions were determined by a vector average method¹⁸ based on direction-of-motion tuning curves in response to gratings. The grating motion direction closest to this vector average was taken as the preferred direction. A neuron was deemed directionally selective if the direction index exceeded 0.5 (direction index = $1 - (\text{anti-preferred activity})/(\text{preferred activity})$; baseline activity subtracted). All cells in this study were directionally selective by this definition.

Cross-correlograms were calculated for each stimulus and prestimulus condition at 1-ms resolution from -50 to $+50$ ms averaged over 10–40 trials. In addition, a peristimulus time histogram (PSTH) predictor was calculated to correct for stimulus-locked synchrony¹⁹. Synchrony was quantified by the NCC¹⁰. We also calculated z-scores^{19,20}, which yielded the same pattern of results as the NCC data (see Supplementary section 7). The NCC was calculated over correlogram time intervals of ± 5 ms, ± 10 ms, ± 20 ms and ± 30 ms relative to time zero. Our overall results were robust with respect to the choice of time intervals. Most pairs had a synchrony peak width of around 10–20 ms centred at time zero. Accordingly, the data plotted in the figures and the corresponding *P*-values were derived from ± 10 -ms time intervals. For an NCC to be included in our data set, the cross-correlograms had to have at least 200 entries. This criterion was usually far exceeded. The CPI was derived from previous work^{11,12} (see Supplementary section 8).

To classify neurons as component- or pattern-like, we computed the average CPI across all three plaid pattern configurations for each neuron. Neurons with large average CPI values (in the upper 50%) of this distribution were classified as component-type whereas neurons with small average CPI values (in the lower 33%) were classified as pattern-type (for justification, see Supplementary section 9). To examine neuronal synchrony unaffected by eye movements, we eliminated periods during and immediately after microsaccades, tracking eye-movements, and eye drifts from our data set. Eye-movement-free periods were subdivided into 200-ms sections and these sections were treated as though they constituted individual trials (see Supplementary section 10).

Received 10 July; accepted 1 November 2002; doi:10.1038/nature01285.

- Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
- Singer, W. & Gray, C. M. Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* **18**, 555–586 (1995).
- von der Malsburg, C. & Schneider, W. A neural cocktail party processor. *Biol. Cybern.* **54**, 29–40 (1986).
- Stoner, G. R. & Albright, T. D. Neural correlates of perceptual motion coherence. *Nature* **358**, 412–414 (1992).
- Stoner, G. R., Albright, T. D. & Ramachandran, V. S. Transparency and coherence in human motion perception. *Nature* **344**, 153–155 (1990).
- Stoner, G. R. & Albright, T. D. Motion coherency rules are form-cue invariant. *Vision Res.* **32**, 465–475 (1992).
- Stoner, G. R. & Albright, T. D. Luminance contrast affects motion coherency in plaid patterns by acting as a depth-from-occlusion cue. *Vision Res.* **38**, 387–401 (1998).
- Lindsey, D. T. & Todd, J. T. On the relative contributions of motion energy and transparency to the perception of moving plaids. *Vision Res.* **36**, 207–222 (1996).
- Singer, W. Neuronal synchrony: A versatile code for the definition of relations. *Neuron* **24**, 49–65 (1999).
- Bair, W., Zohary, E. & Newsome, W. T. Correlated firing in macaque visual area MT: time scales and relationship to behavior. *J. Neurosci.* **21**, 1676–1697 (2001).
- Movshon, J. A., Adelson, E. H., Gizzi, M. & Newsome, W. T. *Study Group on Pattern Recognition* (eds Chagas, T., Gattass, R. & Gross, C. G.) 117–151 (Pontificia Academia Scientiarum, Vatican City, 1985).
- Rodman, H. R. & Albright, T. D. Single-unit analysis of pattern-motion-selective properties in the middle temporal visual area. *Exp. Brain Res.* **75**, 53–64 (1989).
- Castelo-Branco, M., Goebel, R., Neuenschwander, S. & Singer, W. Neural synchrony correlates with surface segregation rules. *Nature* **405**, 685–689 (2000).
- Pack, C. C., Berezovskii, V. K. & Born, R. T. Dynamic properties of neurons in cortical area MT in alert and anaesthetized macaque monkeys. *Nature* **414**, 905–908 (2001).
- de Oliveira, S. C., Thiele, A. & Hoffmann, K. P. Synchronization of neuronal activity during stimulus expectation in a direction discrimination task. *J. Neurosci.* **17**, 9248–9260 (1997).

- Kreiter, A. K. & Singer, W. Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *J. Neurosci.* **16**, 2381–2396 (1996).
- Thiele, A., Distler, C. & Hoffmann, K. P. Decision-related activity in the macaque dorsal visual pathway. *Eur. J. Neurosci.* **11**, 2044–2058 (1999).
- Batschelet, E. *Animal Orientation and Navigation* (eds Galler, S. R., Schmidt-Koenig, K., Jacobs, G. J. & Belleville, R. E.) 61–91 (NASA Aeronautics and Space Administration, Washington DC, 1972).
- Aertsen, A. M., Gerstein, G. L., Habib, M. K. & Palm, G. Dynamics of neuronal firing correlation: modulation of "effective connectivity". *J. Neurophysiol.* **61**, 900–917 (1989).
- Eggermont, J. J. & Smith, G. M. Neural connectivity only accounts for a small part of neural correlation in auditory cortex. *Exp. Brain Res.* **110**, 379–391 (1996).
- Stoner, G. R. & Albright, T. D. The interpretation of visual motion: evidence for surface segmentation mechanisms. *Vision Res.* **36**, 1291–1310 (1996).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank J. Constanza and D. Diep for technical assistance. We also thank T. D. Albright, J. R. Reynolds, G. Boynton and J. Hegde for discussions on this work and comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.S. (e-mail: gene@salk.edu).

Selective gating of visual signals by microstimulation of frontal cortex

Tirin Moore & Katherine M. Armstrong

Department of Psychology, Princeton University, Princeton, New Jersey 08544, USA

Several decades of psychophysical and neurophysiological studies have established that visual signals are enhanced at the locus of attention^{1–5}. What remains a mystery is the mechanism that initiates biases in the strength of visual representations⁶. Recent evidence argues that, during spatial attention, these biases reflect nascent saccadic eye movement commands^{7,8}. We examined the functional interaction of saccade preparation and visual coding by electrically stimulating sites within the frontal eye fields (FEF) and measuring its effect on the activity of neurons in extrastriate visual cortex. Here we show that visual responses in area V4 could be enhanced after brief stimulation of retinotopically corresponding sites within the FEF using currents below that needed to evoke saccades. The magnitude of the enhancement depended on the effectiveness of receptive field stimuli as well as on the presence of competing stimuli outside the receptive field. Stimulation of non-corresponding FEF representations could suppress V4 responses. The results suggest that the gain of visual signals is modified according to the strength of spatially corresponding eye movement commands.

Important stimuli are generally the targets of eye movements, although they can be selectively processed without direct gaze¹. The ability of primates to dissociate the point of gaze from the point of interest does not preclude a common mechanism of attention and eye movement control. Instead, this ability may represent a specialized adaptation that allows covert monitoring, while avoiding the consequences of direct eye contact⁹. For example, it may be adaptive to covertly monitor a surly-looking gentleman seated nearby on a train, whereas confronting him with direct gaze could prove hazardous. Under normal circumstances, shifts in gaze occur freely, and visual detection is facilitated at the location of impending saccades even before the eyes move^{10,11}.

The involvement of the FEF in the selection of visual targets for saccades is well established¹². The FEF contains neurons that have a broad spectrum of properties, from visual to motor, culminating in a map of visual space in which the amplitude and direction of saccades are organized in retinotopic coordinates¹³. Electrical stimulation of the FEF evokes short-latency saccades in both human and non-human primates^{13–15}. Stimulation of the FEF with currents below the movement threshold does not evoke saccades, but nonetheless biases the selection of targets for eye movements^{16,17}. Subthreshold stimulation of eye movement representations within the FEF can also improve a monkey's ability to covertly filter visual stimuli⁸. This result suggests that microstimulation of the FEF not only initiates the selection of a particular eye movement, but also initiates the selection of particular visual stimuli at the location to which the movement is directed. We studied the influence of subthreshold FEF stimulation on the visual responses of neurons in extrastriate area V4. In each of two monkeys, we electrically stimulated sites within the FEF while simultaneously recording the activity of V4 neurons (Fig. 1a). Modulation of activity in area V4 during both overt and covert spatial attention tasks has previously been observed^{5,18}. We reasoned that, if FEF microstimulation initiates both saccade preparation and visual selection, then it should be possible to induce a spatial-attention-like modulation of V4 activity in passively fixating monkeys.

During each experiment, the receptive field (RF) of an isolated V4 neuron was mapped (Fig. 1b, left). With a second electrode, we located a site within the FEF from which saccades could be evoked by electrical stimulation (Fig. 1b, right). Next, we mapped the point in space to which the monkey's gaze was shifted, and measured the current threshold to evoke saccades. At a particular FEF site, the end point of the evoked-saccade vector could fall either within or outside the V4 RF. Finally, using subthreshold currents, we examined the effect of stimulating the FEF site on the visual responses of the V4 neuron. Stimulation was applied to the FEF site 200–500 ms after the appearance of visual stimuli. This delay allowed us to examine the effects of FEF stimulation on visually driven activity in

V4. Thus, stimulation could amplify, have no effect on, or interfere with the V4 representation of the stable stimulus.

An example of the effect of FEF stimulation on the response of a V4 neuron is shown in Fig. 2. Stimulation of the FEF site evoked saccades to a location 11° into the lower contralateral quadrant and into the V4 RF. We applied 50-ms trains of subthreshold current pulses (20 μ A) to the FEF site during the presentation of stable oriented-bar stimuli while the monkey maintained central fixation. Responses of the neuron during FEF stimulation were excluded from the analysis to avoid contamination by the stimulation artefact. This neuron responded well to the appearance of an optimally oriented bar stimulus in its RF, but adapted within a few hundred milliseconds (Fig. 2a). However, after FEF stimulation, the response was transiently enhanced compared with trials in which there was no stimulation. By contrast, stimulation did not affect the activity of the cell when there was no RF stimulus (Fig. 2b). Therefore, the effect of stimulation was dependent on the presence of the RF stimulus.

Figure 2c shows the effect of FEF stimulation as a function of the visually driven activity when no stimulus, the non-preferred, or the preferred stimulus was presented in the RF. We measured the difference in activity between stimulation and control trials during a 70-ms time window immediately after the end of the stimulation train. In this case, FEF stimulation caused a significant enhancement of the firing rate of the V4 cell when there was a preferred visual stimulus in its RF (mean response change, 22.1 spikes s^{-1} ; paired *t*-test, *P* < 0.05). By contrast, there was no effect of FEF stimulation when the cell was not driven by a RF stimulus (mean

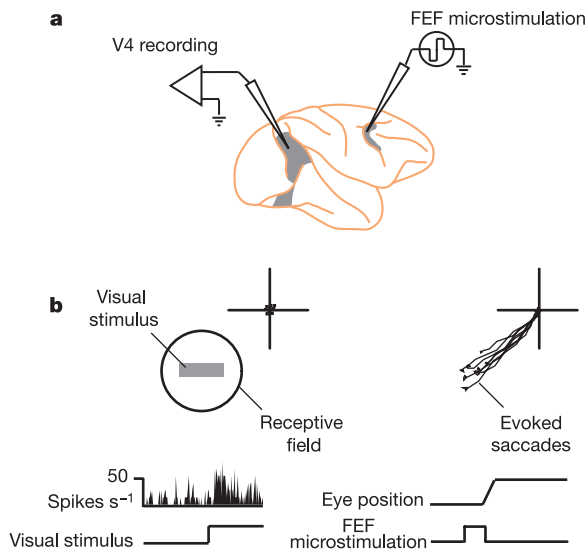


Figure 1 Mapping receptive fields and evoking saccades. **a**, Sites within the frontal eye fields (FEF) were electrically stimulated while recording from neurons in area V4. The cartoon shows a side view of the macaque brain. The locations of the FEF in the arcuate sulcus and of area V4 in the prelunate gyrus are shown (shaded). **b**, Left, the response of a V4 neuron when a visual stimulus appears in its receptive field (RF) while the monkey maintains central fixation (dots at origin). Right, saccades evoked to a location spatially corresponding to the V4 RF following suprathreshold microstimulation of an FEF site.

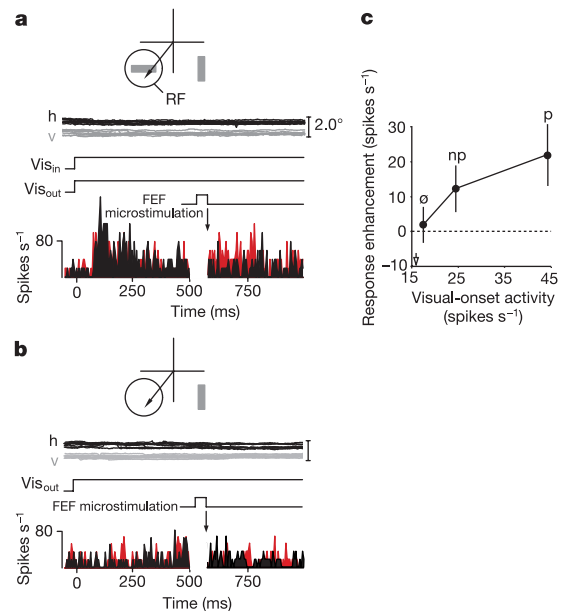


Figure 2 Effect of FEF stimulation on the response of a V4 neuron. The neuron's RF (circle) is aligned with the saccade vector (arrow) evoked with suprathreshold FEF stimulation (>40 μ A). **a**, Histograms show the V4 neuron's mean response during control (black) and stimulation (red) conditions (ten trials per condition, 5-ms bins). The control histogram is shown superimposed on the stimulation histogram for two time periods: after the onset of RF (Vis_{in}) and non-RF (Vis_{out}) stimuli and after a 50-ms subthreshold (20 μ A) stimulation train had been applied to the FEF site. Above, black (horizontal; h) and grey (vertical; v) traces show the monkey's eye position. **b**, As in **a**, except that histograms show the neuron's response when the RF stimulus was not presented. **c**, Response enhancement plotted as a function of visual-onset activity. The x-axis shows the average activity during a 400-ms time window after the onset of the visual display, with 0 (no stimulus), np (a non-preferred stimulus) or p (a preferred stimulus) in the RF. The y-axis shows the mean change in activity (stimulation – control) for a 70-ms time window following the offset of FEF stimulation. Arrow indicates the baseline activity. Error bars represent the s.e.m.

response change, $1.9 \text{ spikes s}^{-1}$; paired t -test, $P > 0.70$). During the presentation of a bar stimulus at a non-preferred orientation, FEF stimulation resulted in an intermediate enhancement (mean response change, $12.3 \text{ spikes s}^{-1}$; paired t -test, $P > 0.05$). Thus, the effect of stimulation seemed to depend on how well the V4 neuron was visually driven.

We tested the effect of FEF microstimulation on a total of 64 V4 neurons in the two monkeys. Of these, 31 neurons had RFs into which the evoked saccade shifted the monkey's gaze, whereas 28 had RFs that the saccade end points fell outside and were displaced from by an angle of $>45^\circ$. Two further neurons were tested in both conditions by advancing the FEF electrode to a new site between experiments while recording from the same V4 cell. The remaining three neurons had RFs that partially overlapped the FEF representation and/or were tested with visual stimuli that bridged the two representations, and therefore were excluded from the analysis. The FEF was always stimulated with currents below the site's movement threshold current. The mean test current ($40.7 \mu\text{A}$) was 55% of the mean threshold current ($73.5 \mu\text{A}$). During each experiment, optimal and non-optimal visual stimuli were presented within the V4 RF, outside it, or at both locations simultaneously. These conditions allowed us to examine whether the stimulation effect depended on how well the RF stimulus drove the V4 cell, and whether it depended on the presence or absence of a potentially distracting stimulus outside the RF.

When the evoked saccade shifted the monkey's gaze into the V4 RF, subthreshold FEF stimulation resulted in a visually dependent enhancement of the population response (Fig. 3a). In this analysis, data were collapsed across the distracter conditions. When a stimulus was presented in the RF, stimulation enhanced V4 responses (paired t -test, $P < 0.0001$). By contrast, FEF stimulation had no effect on the response of the population when there was no RF stimulus (paired t -test, $P > 0.25$). The dependence of the stimulation effect on the visually driven activity was also evident when we excluded the conditions with no RF stimulus. That is, the enhancement depended on whether the preferred or non-preferred stimulus appeared in the RF, the preferred stimulus yielding greater response enhancement (analysis of variance (ANOVA), $P < 0.02$). Thus, the enhancement grew with increasing visual drive.

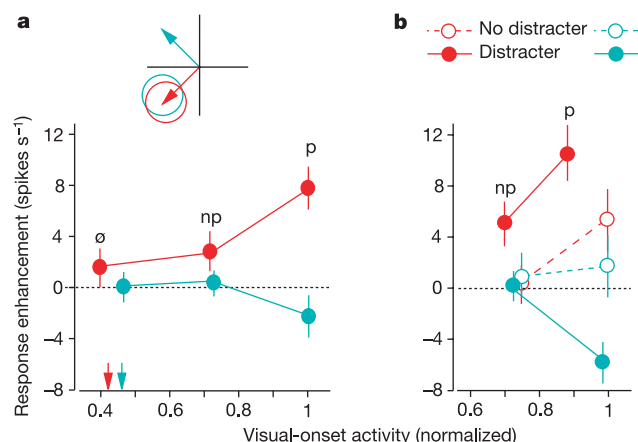


Figure 3 Effect of FEF stimulation on the responses of populations of V4 neurons. Data were obtained from experiments in which the evoked saccades (arrows) shifted the monkey's gaze to a point within (red, $n = 33$) or outside (blue, $n = 30$) the RF (circles) of the recorded V4 neuron. **a**, The mean response difference between stimulation and control conditions (stimulation — control) plotted as a function of visual-onset activity. The onset activity of each neuron in the population was normalized to its response to the preferred stimulus presented alone. Arrows indicate the mean baseline activity for each population subset. **b**, Same as **a**, but conditions with and without a non-RF stimulus (distracter) are shown separately.

We also examined the dependence of the stimulation effect on the presence or absence of a distracter stimulus outside the RF. We compared the conditions in which the RF stimulus was presented either alone or simultaneously with the distracter stimulus (Fig. 3b). The effect of FEF stimulation was greater when a non-RF stimulus was presented simultaneously with the RF stimulus than when the RF stimulus was presented in isolation (ANOVA, $P < 0.02$). The effects of distracter and RF stimuli did not significantly interact (ANOVA, $P > 0.5$). As a result, we observed the largest response enhancement when the preferred stimulus was presented in the RF simultaneously with a distracter outside the RF. In this condition, FEF stimulation increased the response of V4 neurons by an average of $10.4 \text{ spikes s}^{-1}$. The population response to the onset of the preferred stimulus averaged 48 spikes s^{-1} . Thus, stimulation enhanced visual responses by more than 20% of the maximum.

The effect of subthreshold stimulation was different when the evoked saccade shifted the monkey's gaze to a location outside the V4 RF. In these experiments, the distracter stimulus was placed at the end point of the evoked-saccade vector. When the preferred visual stimulus was presented in the RF simultaneously with the non-RF distracter, the condition that produced the largest enhancement during stimulation of overlapping representations, we observed a suppression of V4 responses (paired t -test, $P < 0.01$) (Fig. 3b). Following FEF stimulation, the population response was reduced by an average of $5.1 \text{ spikes s}^{-1}$. In all other conditions, FEF stimulation did not alter V4 responses (Fig. 3b). Thus, FEF stimulation could enhance or suppress the visual response depending on whether the two cortical representations overlapped retinotopically.

Microstimulation of the FEF could have altered the activity of V4 neurons in a number of ways. Stimulation could have directly activated V4 neurons antidromically, given that at least some cells in V4 project to the FEF¹⁹. However, our failure to change the activity of these neurons when they were not driven by a RF stimulus indicates that direct antidromic effects were absent. Furthermore, the dependence of stimulation effects on the presence of a RF stimulus indicates that activation of the FEF site did not impose an extraneous visual signal on V4 neurons. This interpretation is consistent with the failure of human subjects to report visual sensations during cortical stimulation of their homologous FEF^{14,15}. Second, although we stimulated the FEF with subthreshold currents, we may have nonetheless destabilized the monkey's fixation and increased the frequency of microsaccades. Such an increase could have raised the visual responsiveness of each neuron by effectively moving the stimulus within the RF²⁰. However, this effect would not depend on the overlap between the FEF and V4 representations, contrary to our observations.

Instead, microstimulation of FEF sites appears to have activated a network that controls the gain of visually driven signals. Moreover, our results show that activation of this network biases not only the selection of eye movements^{16,17}, but also the strength of visual cortical signals, revealing a common network for the control of visual and oculomotor selection. In addition, the dependence of the stimulation effects on the features of the RF stimulus as well as on the presence of a non-RF distracter indicates that the above network may be involved in previously reported attentional modulations in area V4. Studies in which changes in V4 responses are observed during spatial attention tasks have shown that the magnitude of modulation depends on how well the RF stimulus drives the V4 cell²¹ and on the number of distracter stimuli⁵. In our study, the display conditions in which FEF stimulation produced the largest effect were similar to those that gave rise to the largest V4 modulation in monkeys trained to direct attention either inside or outside a V4 RF. It is also noteworthy that, under these conditions, the magnitude of the enhancement effect ($\sim 20\%$) was comparable to that previously observed in V4 during spatial-attention tasks²¹. Although it is tempting to speculate that the

observed effects resulted from orthodromic activation of FEF neurons with top-down projections to posterior visual areas²², including direct projections to V4 as well as to parietal and temporal areas, this is only one of many possible ways in which stimulation could gate signals in area V4. Future experiments should aim to specify further the pathway or pathways that are sufficient to bring about the observed gain modulation.

Primate vision consists of a sequential sampling of the details contained within a scene²³. The extraction of information from a scene depends on the continual shifting of perceptual resources, such as the foveas, across locations in space. This process must include the reciprocal interaction of brain mechanisms that are involved primarily in coding the visual stimulus with those that are involved primarily in moving the eye²⁴. The results of the current study reveal an equivalence of biases in oculomotor preparation and biases in the gain of spatially corresponding visual signals when oculomotor plans are not carried out. As such, covert spatial attention may simply reflect one emergent property of visuomotor integration. □

Methods

We recorded the responses of V4 cells to oriented-bar stimuli while the monkey fixated a central spot and while subthreshold stimulation trains lasting 20–50 ms (200 Hz, biphasic, 0.2-ms pulse duration) were applied to the FEF. We determined the metrics of the evoked saccade and the movement threshold of each FEF site in a separate behavioural paradigm in which stimulation with varying current amplitude was delivered to the site during the performance of a fixation task⁸. During the experiment, visual stimuli (50–93% Michelson contrast) were presented for 1.5–2 s inside and outside the RF of a V4 neuron. RFs were mapped in a further behavioural paradigm in which oriented bars were swept across the display during fixation while monitoring the activity of the recorded cell²⁵. Bar stimuli were presented either at the preferred orientation or orthogonal to the preferred orientation during successive trials. All visual stimuli were displayed on a video monitor (30 cm vertical × 40 cm horizontal, 60 Hz) positioned 57 cm in front of the monkey. Throughout behavioural testing, eye position was monitored through a scleral search coil and stored at 200 Hz. All general surgical and experimental procedures, which have previously been described²⁶, were approved by the Princeton University Animal Care and Use Committee and were in accordance with National Institutes of Health guidelines.

Received 15 October; accepted 25 November 2002; doi:10.1038/nature01341.

1. Sperling, G. & Melchner, M. J. The attention operating characteristic: examples from visual search. *Science* **202**, 315–318 (1978).
2. Blaser, E., Sperling, G. & Lu, Z. L. Measuring the amplification of attention. *Proc. Natl Acad. Sci. USA* **96**, 11681–11686 (1999).
3. Moran, J. H. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
4. Haenny, P. E. & Schiller, P. H. State dependent activity in monkey visual cortex. I. Single cell activity in V1 and V4 on visual tasks. *Exp. Brain Res.* **69**, 225–244 (1988).
5. Motter, B. C. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70**, 909–919 (1993).
6. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
7. Rizzolatti, G., Riggio, L., Dascola, I. & Umiltà, C. Reorienting attention across the horizontal and vertical meridians: evidence in favor of premotor theory of attention. *Neuropsychologia* **25**, 31–40 (1987).
8. Moore, T. & Fallah, M. Control of eye movements and spatial attention. *Proc. Natl Acad. Sci. USA* **98**, 1273–1276 (2001).
9. van Hooff, J. A. R. A. M. in *Non-Verbal Communication* Ch. 7 (ed. Hinde, R. A.) Ch. 7 (Cambridge Univ. Press, Cambridge, UK, 1972).
10. Shepherd, M., Findlay, J. M. & Hockey, R. J. The relationship between eye movements and spatial attention. *Q. J. Exp. Psychol.* **A 38**, 475–491 (1986).
11. Hoffman, J. E. & Subramaniam, B. The role of visual attention in saccadic eye movements. *Percept. Psychophys.* **57**, 787–795 (1995).
12. Tehovnik, E. J., Sommer, M. A., Chou, I. H., Slocum, W. M. & Schiller, P. H. Eye fields in the frontal lobes of primates. *Brain Res. Brain Res. Rev.* **32**, 413–448 (2000).
13. Bruce, C. J., Goldberg, M. E., Bushnell, M. C. & Stanton, G. B. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J. Neurophysiol.* **54**, 714–734 (1985).
14. Penfield, W. & Rasmussen, T. *The Cerebral Cortex of Man: A Clinical Study of Localization of Function* (Macmillan, New York, 1950).
15. Godoy, J., Lunders, H., Dinner, D. S., Morris, H. H. & Wyllie, E. Versive eye movements elicited by cortical stimulation of the human brain. *Neurology* **40**, 296–299 (1990).
16. Schiller, P. H. & Tehovnik, E. J. Look and see: how the brain moves your eyes about. *Prog. Brain Res.* **134**, 127–142 (2001).
17. Burman, D. D. & Bruce, C. J. Suppression of task-related saccades by electrical stimulation in the primate's frontal eye field. *J. Neurophysiol.* **77**, 2252–2267 (1997).
18. Moore, T., Tolia, A. S. & Schiller, P. H. Visual representations during saccadic eye movements. *Proc. Natl Acad. Sci. USA* **95**, 8981–8984 (1998).

19. Schall, J. D., Morel, A., King, D. J. & Bullier, J. Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. *J. Neurosci.* **15**, 4464–4487 (1995).
20. Leopold, D. A. & Logothetis, N. K. Microsaccades differentially modulate neural activity in the striate and extrastriate visual cortex. *Exp. Brain Res.* **123**, 341–345 (1998).
21. McAdams, C. J. & Maunsell, J. H. R. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* **19**, 431–441 (1999).
22. Stanton, G. B., Bruce, C. J. & Goldberg, M. E. Topography of projections to posterior cortical areas from the macaque frontal eye fields. *J. Comp. Neurol.* **353**, 291–305 (1995).
23. Yarbus, A. L. *Eye Movements and Vision* (Plenum, New York, 1967).
24. Moore, T. Shape representations and visual guidance of saccadic eye movements. *Science* **285**, 1914–1917 (1999).
25. Zipser, K., Lamme, V. A. F. & Schiller, P. H. Contextual modulation in primary visual cortex. *J. Neurosci.* **16**, 7376–7389 (1996).
26. Nakamura, K., Chung, H. H., Graziano, M. S. & Gross, C. G. Dynamic representation of eye position in the parieto-occipital sulcus. *J. Neurophysiol.* **81**, 2374–2385 (1999).

Acknowledgements We thank C. G. Gross and M. Fallah for their help. This work was supported by grants from the National Institutes of Health, and a HHMI predoctoral fellowship to K.M.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T. M. (e-mail: tirin@princeton.edu).

Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders

Ivelisse Sánchez[†], Christian Mahlike^{*} & Junying Yuan^{*}

^{*} Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

The expansion of a CAG repeat coding for polyglutamine in otherwise unrelated gene products is central to eight neurodegenerative disorders including Huntington's disease¹. It has been well documented that expanded polyglutamine fragments, cleaved from their respective full-length proteins, form microscopically visible aggregates in affected individuals and in transgenic mice^{2–7}. The contribution of polyglutamine oligomers to neurodegeneration, however, is controversial. The azo-dye Congo red binds preferentially to β -sheets containing amyloid fibrils^{8,9} and can specifically inhibit oligomerization¹⁰ and disrupt preformed oligomers. Here we show that inhibition of polyglutamine oligomerization by Congo red prevents ATP depletion and caspase activation, preserves normal cellular protein synthesis and degradation functions, and promotes the clearance of expanded polyglutamine repeats *in vivo* and *in vitro*. Infusion of Congo red into a transgenic mouse model of Huntington's disease, well after the onset of symptoms, promotes the clearance of expanded repeats *in vivo* and exerts marked protective effects on survival, weight loss and motor function. We conclude that oligomerization is a crucial determinant in the biochemical properties of expanded polyglutamine that are central to their chronic cytotoxicity.

To assess directly the role of oligomerization in the cytotoxicity of expanded polyglutamines, we determined whether anti-amyloid compounds that are known to bind to β -sheet-containing amyloid fibrils, including polyanions¹¹, tetracyclines¹² and azo-dyes such as Congo red^{8,10} and chrysamine G¹³, could prevent expanded poly-

[†] Present address: Department of Anatomy and Neurobiology, Boston University School of Medicine, 715 Albany Street, L-813, Boston, Massachusetts 02118, USA.

glutamine induced cell death (Fig. 1). We transfected HeLa cells with an expression construct encoding an expanded polyglutamine tract tagged with haemagglutinin A (HA-Q79). As shown previously, the expression of Q79 resulted in cell death (Fig. 1a), detectable 48 h after transfection¹⁴. Congo red, but none of the other agents tested, showed a remarkable protective effect (60%) against cell death induced by HA-Q79 (Fig. 1a). Given these results, we examined the specific effects and mechanism by which the β -sheet-binding compound Congo red exerts protection against polyglutamine-induced toxicity.

Changes in energy metabolism have been detected in brain tissue

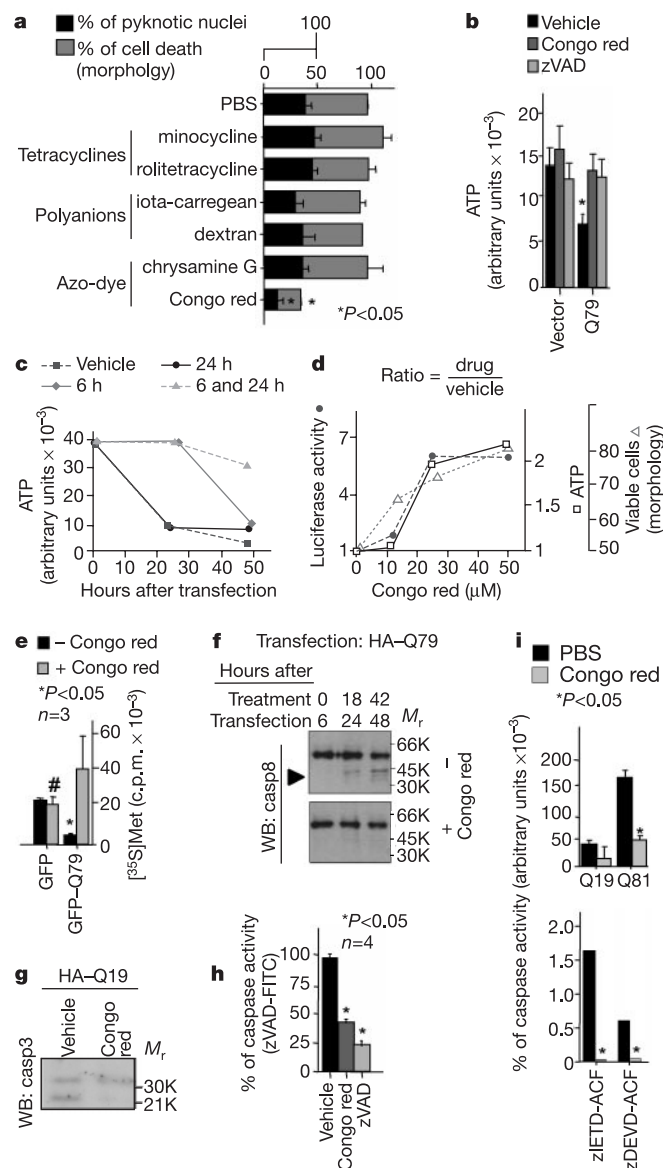


Figure 1 Effect of anti-amyloid compounds on cell death induced by polyglutamine. **a**, Percentage of cell death in HeLa cells expressing HA-Q79 treated with the indicated compounds (100 μ M) for 42 h starting at 6 h after transfection. **b**, **c**, ATP concentrations in control or HA-Q79-transfected HeLa cells and in cells treated with 100 μ M Congo red (100 μ M), zVAD (100 μ M) or PBS for 42 hrs (**b**) or at the indicated number of hours after transfection (**c**). **d**, Dose-dependent effects of Congo red. **e–i**, Effects of Congo red on protein synthesis assessed by [³⁵S]methionine labelling (**e**), on activation of caspase-8 and caspase-3 (**f**, **g**), and on caspase activity in cells using a zVAD-FITC assay (**h**), in a cell-free system using zIETD.AMC (**i**, top) or in cells lysates using zIETD.AMC or zDEVD.AMC (**i**, bottom).

from presymptomatic animal models of, and presymptomatic individuals with, Huntington's disease¹⁵. We detected a sharp decrease in concentrations of ATP on expression of expanded polyglutamine in cells, which reached close to minimal levels as early as 24 h after transfection (Fig. 1b–d). This ATP depletion, however, was inhibited by the addition of zVAD.fmk, a polycaspase inhibitor, suggesting that changes in the energy metabolism leading to the reduction in ATP take place downstream of caspase activation (Fig. 1b). Congo red significantly inhibited the loss of ATP induced by expression of expanded polyglutamine in a time- and dose-dependent manner (Fig. 1c, d). Addition of Congo red 6 h after transfection prevented ATP depletion. A reduced but detectable protective effect was also achieved when Congo red was added as late as 24 h after transfection, a time when aggregated, expanded polyglutamine was already visible in untreated control cells (Fig. 1c). An additive effect of two doses (6 and 24 h) of Congo red on the preservation of ATP concentrations was seen (Fig. 1c). No general increases in ATP were detected in control cells treated with Congo red (Fig. 1b).

Inhibition of general protein synthesis is an important early event in cell death¹⁶. To determine whether Congo red inhibits the loss of protein synthesis in cells expressing polyglutamine, we co-transfected HeLa cells with a luciferase construct and HA-Q79, and determined protein synthesis by luciferase activity 48 h after transfection. Expression of HA-Q79 led to a significant loss of luciferase expression. Treatment with Congo red prevented the reduction in luciferase activity, as well as ATP loss and cell morphological changes, in a similar dose-dependent manner (Fig. 1d).

To verify that Congo red exerts a protective effect on general protein synthesis in Q79-expressing cells, we determined the amounts of newly synthesized protein by a [³⁵S]methionine pulse-chase experiment. In agreement with data from the luciferase assay, treatment with Congo red prevented the decrease in newly synthesized ³⁵S-labelled proteins in Q79-expressing cells specifically, but had no effect on the basal amounts of protein synthesis in cells expressing green fluorescent protein (GFP) or in control primary neurons (Fig. 1e and Supplementary Fig. 1a, b). These results indicate that Congo red does not function as a general ATP or protein synthesis enhancer but exerts specific effects upstream to inhibit the loss of cellular functions.

As expanded polyglutamine repeats can recruit and activate caspase-8 (ref. 14), we tested whether treatment with Congo red inhibits the activation of caspase-8 as well as a downstream caspase, caspase-3, in Q79-expressing cells (Fig. 1f, g). HeLa cells were transfected with HA-Q79 in the presence or absence of Congo red, and the cell lysates were analysed by western blot. The appearance of an active caspase-8 fragment of relative molecular mass 45,000 (M_r 45K) was detected at 24 h, preceding the extensive morphological changes but concurrent with the timing of ATP loss (Fig. 1f and Supplementary Fig. 2). Addition of Congo red completely inhibited the appearance of active caspase-8 induced by Q79 expression and consequently also the appearance of activated caspase-3 (Fig. 1g). The inhibition of Q79-induced caspase activation by Congo red was also shown by a fluorogenic general caspase substrate: addition of Congo red or caspase inhibitor zVAD.fmk prevented the appearance of zVAD-FITC-positive cells (Fig. 1h). Consistent with this result, lysates from HA-Q79 expressing cells that had been cultured in the presence of Congo red showed considerably less caspase-8 and -3 activity than did samples treated with PBS (Fig. 1i, bottom).

To determine whether Congo red could inhibit the ability of expanded polyglutamine aggregates to activate caspases independently of cellular metabolic activity, we established a cell-free system with recombinant glutathione S-transferase (GST)-tagged Q19 or Q81 in the presence of HeLa cell S100 lysate and IETD.AFC, a fluorogenic substrate for caspase-8 like activity (Fig. 1i, top). The presence of GST-Q81 but not GST-Q19 in HeLa S100 lysate

induced the activation of caspase-8 like activity. Notably, Congo red inhibited the activation of caspase activity in this cell-free system in which the contribution of the proteasome was ruled out by the use of MG132. This experiment suggests that Congo red can inhibit the activation of caspases induced by expanded polyglutamine independently of cell metabolic activity. Taken together, these experiments show that Congo red can inhibit caspase activation induced by expanded polyglutamine in both cells and a cell-free system, and provide biochemical evidence supporting the proposal that Congo red can alter the biochemical and biophysical properties of expanded polyglutamine that underpin its cytotoxicity.

Because overexpression of heatshock chaperone proteins such as HSP40 and HSP70 has been shown to inhibit cell death^{17,18} and seems to alter the properties of polyglutamine oligomers^{19,20}, we determined whether Congo red induced an increase in the expression of endogenous chaperones. Congo red treatment of cells expressing HA-Q79 had no effect on the expression of HSP40 or HSP70 (Supplementary Fig. 2). Thus, Congo red is unlikely to act indirectly through induction of chaperone proteins. In addition, a possible direct inhibitory effect of Congo red on the cell death machinery was ruled out as Congo red did not inhibit any of the pro-apoptotic and necrotic stimuli tested (Supplementary Fig. 3).

Although Congo red did not block the general cell death machinery, it showed an obvious inhibitory effect on the aggregation of expanded polyglutamine, concurrent with its inhibitory effect on cell death (Fig. 2). When SH-SY5Y neuroblastoma cells were treated with Congo red 6 h after transfection with Q79 tagged with GFP (Q79-GFP), the formation of expanded polyglutamine aggregates was reduced significantly (Fig. 2a). In addition,

incubation of the semipurified polyglutamine aggregates with Congo red but not with any of the other compounds resulted in the dissolution of preformed aggregates (Fig. 2b). Thus, Congo red can inhibit nucleation as well as disrupt preformed expanded polyglutamine aggregates.

We designed a chemical absorption assay that measures the remaining compound in the supernatant after incubation with GST-polyglutamine beads and their subsequent removal by centrifugation. The amyloid-like conformation of GST-Q81 and GST-Q62 was confirmed by fluorescence microscopy after staining with Thioflavin T (data not shown). Congo red had a significantly higher affinity to expanded polyglutamine (Q81 and Q62) than to polyglutamine with 19 repeats (Q19) at equal protein concentrations (Fig. 2c), suggesting that Congo red does not bind to polyglutamine *per se*, but interacts with the specific conformation of expanded polyglutamine. Thus, Congo red can differentiate between expanded and normal-length polyglutamine.

To confirm further the ability of Congo red to inhibit polyglutamine oligomerization and to disrupt preformed aggregates, we analysed the state of polyglutamine aggregation by examining equal aliquots of total lysates from cells expressing HA-Q79 that had been treated with Congo red or vehicle, or lysates that were treated with Congo red or vehicle after cell lysis. Cell lysates were passed through a 0.2- μ m acetate filter membrane and the Q79 remaining on the filter membrane was visualized after immunostaining with antibodies against HA (Fig. 2d). HA immunoreactivity for Q79 on the filters was undetectable from lysates isolated from cells treated with Congo red for 48 h starting at 6 h after transfection or from lysates treated with Congo red for 30 min after cell lysis (Fig. 2d). This indicates that Congo red can both inhibit the formation of poly-

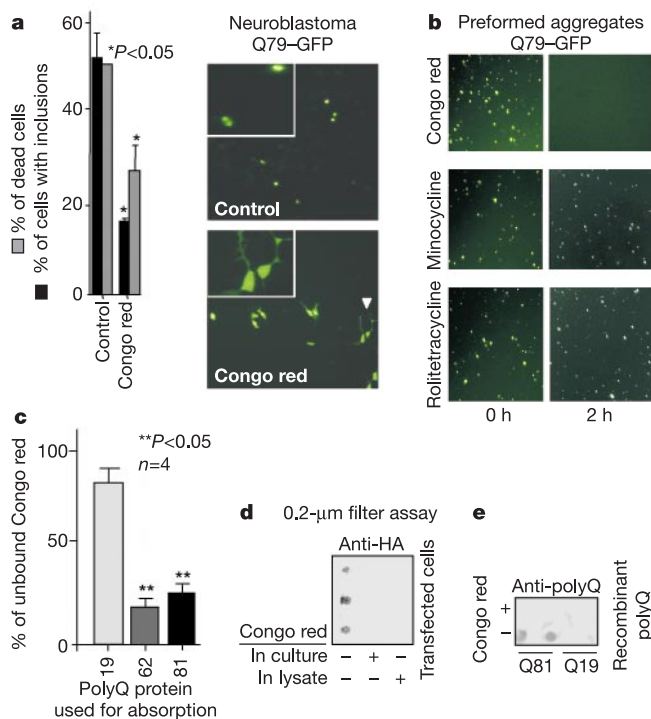


Figure 2 Selective inhibition of polyglutamine oligomerization and cytotoxicity by Congo red. **a**, Cell death and polyglutamine oligomerization in SH-SY5Y cells expressing Q79-GFP treated with 100 μ M Congo red or PBS for 42 h starting at 6 h after transfection. **b**, Lysates from Q79-expressing cells were treated with 100 μ M of indicated compounds (2 h) and visualized by fluorescence microscopy. **c**, Preferential binding of Congo red to Q81 and Q62 rather than Q19, as determined by chemical absorption assay (* $P < 0.01$). **d**, **e**, Effect of Congo red on preformed oligomers (**d**) or on purified recombinant GST-Q19 or GST-Q81 as determined by the filter assay using EM48 antibodies (**e**).

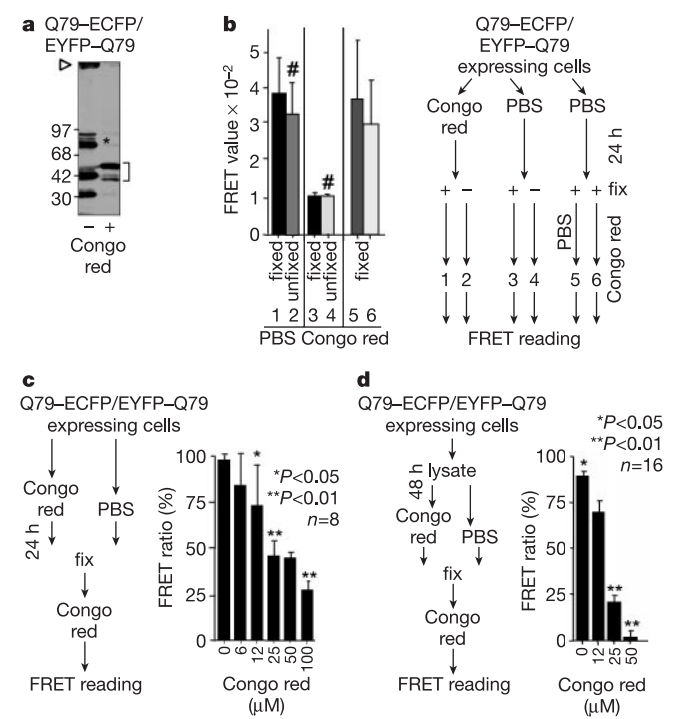


Figure 3 FRET assay for polyglutamine oligomerization. **a**, Western blot of lysates from cells expressing EYFP-Q79 and Q79-ECFP and treated with Congo red (+) or PBS (-) were probed with antibodies against ECFP and EYFP. Open arrowhead indicates the insoluble aggregates at the top of gel. **b**, No significant difference on FRET value occurred after fixation with 1% PFA. **c**, **d**, Dose-dependent effect of Congo red on formation of Q79 aggregates in cells (**c**) and disruption of preformed oligomers from cell lysates (**d**), as determined by FRET.

glutamine aggregates (at least larger than 0.2 μm in size) in cells and dissolve preformed polyglutamine aggregates.

We confirmed that Congo red can inhibit aggregate formation in the absence of other cellular proteins by carrying out the filter assay with bacterially expressed and purified recombinant Q81 and Q19 incubated with either vehicle or 100 μM Congo red (Fig. 2e). In contrast to Congo red, compounds such as chrysamine G, which is structurally very similar to Congo red and can bind expanded polyglutamine aggregates (data not shown), did not disrupt the formation, dissolve preformed polyglutamine aggregates or inhibit cell death induced by expanded polyglutamine in our assays (Fig. 1). These results indicate that the ability of Congo red to inhibit and disrupt polyglutamine oligomerization but not binding *per se* is crucial to the inhibition of cell death induced by expanded polyglutamine repeats.

To rule out the possibility that Congo red merely disrupts the large polyglutamine aggregates without affecting the formation of oligomeric polyglutamine units, which may be indistinguishable by conventional fluorescent microscopy, we used a fluorescence resonance energy transfer (FRET) Q79 assay. Enhanced yellow fluorescent protein (EYFP)-tagged Q79 and enhanced cyan fluorescent protein (ECFP)-tagged Q79 were co-transfected into HeLa cells (Fig. 3a). On western blot with antibodies against EYFP/ECFP, lysates from cells transfected with HA-Q79-ECFP and EYFP-HA-Q79 but untreated with Congo red showed several high M_r bands including SDS-insoluble HA-Q79-ECFP and EYFP-HA-Q79 aggregates unable to enter the separating gel (Fig. 3a, arrowhead) and oligomers that entered the gel (Fig. 3a, asterisk). By contrast, lysates from cells treated with Congo red contained two protein bands close to the predicted M_r for HA-Q79-ECFP and EYFP-HA-Q79 on SDS polyacrylamide gel electrophoresis (SDS-PAGE; ref. 21 and Fig. 3a, brackets).

The oligomeric interaction of expanded polyglutamine in cells was examined further by FRET between HA-Q79-ECFP and EYFP-HA-Q79 24 h after their transfection into cells cultured in the presence or absence of Congo red (Fig. 3b–d). The value of FRET was determined as the ratio between fluorescence at 535 nm and

that at 460 nm after excitation at 430 nm. To exclude the effect of Congo red absorption, we defined the ratio of FRET as the value of FRET in cells treated with Congo red divided by the maximum FRET in control cells to which Congo red was added after fixation. To determine the maximum value of FRET, we co-transfected cells with HA-Q79-ECFP and EYFP-HA-Q79, and 24 h after transfection we fixed them with 1% paraformaldehyde. Congo red was added after fixation and the value of FRET was determined as mentioned above and used as a 100% control value for each concentration of Congo red used. Fixation by 1% paraformaldehyde *per se* had no effect on FRET (Fig. 3b). Cells cultured in the presence of Congo red (6–50 μM) were also fixed with 1% paraformaldehyde 24 h after transfection, and Congo red was then added to the original treatment concentration to normalize for absorption effects. Treatment with Congo red caused a dose-dependent inhibition of the FRET ratio between Q79-ECFP and EYFP-Q79, indicating that Congo red can inhibit Q79 oligomerization (Fig. 3c).

To verify the ability of Congo red to disrupt preformed oligomers independently of cellular function, we determined the effect of Congo red on isolated polyglutamine aggregates by FRET (Fig. 3d). Cells transfected with HA-Q79-ECFP and EYFP-HA-Q79 were lysed, treated with Congo red or PBS, and fixed with 1% paraformaldehyde. The FRET value and ratio were determined as above in the presence of Congo red after fixation. Preformed Q79 oligomers were also disrupted by Congo red, as indicated by the dose-dependent decreases in FRET after Congo red treatment (Fig. 3d). These experiments confirm that Congo red can inhibit the formation of, as well as disrupt, preformed polyglutamine oligomeric units independently of cellular function.

To examine whether disruption of polyglutamine oligomerization interferes with its interaction and the concurrent aggregation of other proteins, we co-transfected cells with Q79 and one of the following expression constructs, GFP-FADDdn, GFP-caspase-8dn or GFP-tagged to exon 1 of the *huntingtin* gene (HDQ25-GFP), in the presence of Congo red or zVAD.fmk. Congo red but not zVAD.fmk clearly prevented the recruitment and aggregation of GFP-FADDdn, GFP-caspase-8dn and HDQ25-GFP (Fig. 4a, b). Congo red treatment also prevented the recruitment of endogenous FADD, or caspase-8 by expanded polyglutamine aggregates, as shown by the 0.2- μm filter assay (Fig. 4c). These results underscore the selectivity of Congo red on the oligomeric expanded polyglutamine repeats and its inhibitory effects on the subsequent downstream events, including aberrant protein–protein interactions and aggregation.

Because these expanded polyglutamine oligomers accumulate in cells and seem to have a significantly slower turnover rate than that of shorter polyglutamine²², we tested the effect of Congo red on the turnover of Q79 oligomers by measuring the amounts of [³⁵S]methionine pulse-labelled polyglutamine. The quantities of newly synthesized Q79 1 h after chase were higher in cells treated with Congo red than in untreated cells, clearly showing that Congo red does not suppress the expression of Q79 (Fig. 5a, top). But the amounts of [³⁵S]-labelled HA-Q79 were over twofold less after 24 h of pulse chase in cells treated with Congo red, suggesting that there is a faster turnover rate of HA-Q79 in the presence of Congo red than in its absence (Fig. 5a, top, and 5b). In line with these results, the steady-state quantities of Q79 were more than twofold less after 24 h in the presence of Congo red than in its absence (Fig. 5a, bottom).

As we had already determined that Congo red does not function as a general enhancer of protein synthesis (Fig. 1), we wanted to rule out the possibility that Congo red may affect general protein degradation by testing its effect on proteasome-mediated protein degradation. Whereas protein degradation was inhibited by the general proteasome inhibitor MG132, Congo red or zVAD had no effect on general protein degradation (Supplementary Fig. 4). To test whether inhibition of Q79 aggregation also prevents the reported inhibition in proteasome activity^{23,24}, we measured the

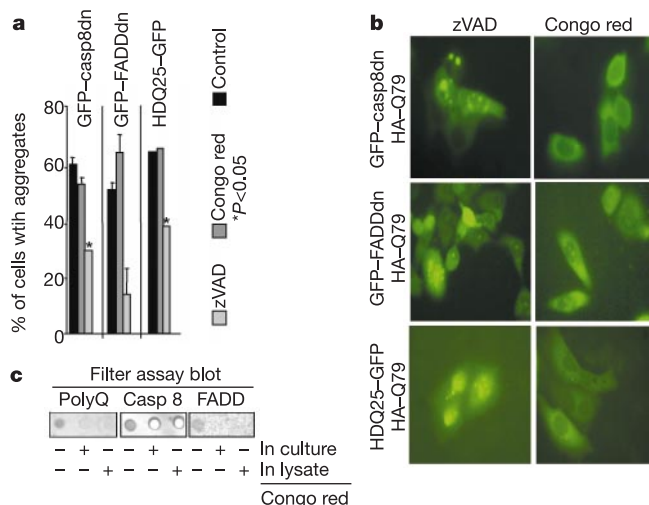


Figure 4 Inhibition of polyglutamine oligomerization prevents abnormal protein recruitment. **a**, **b**, Effect of Congo red on the recruitment by Q79 of GFP-tagged caspase-8dn, FADDdn and huntingtin Q25 in the HeLa cells. Cells co-transfected with HA-Q79 and one of the indicated constructs above were pretreated with 100 μM zVAD.fmk or 100 μM Congo red. **c**, Lysates from cells expressing HA-Q79 treated with 100 μM Congo red or PBS alone in culture or after lysis were passed through a 0.2- μm filter. The presence of expanded polyglutamine aggregates, endogenous caspase-8 and FADD in the filter was detected by EM48 antibodies, antibodies against caspase-8 or antibodies against FADD, respectively.

a Chase (h) 1 24
Congo red - - + +
Pulse labelled
46
30
21
Steady state
46
30
21
Tranf: HA-Q79
IP: anti-HA
WB: anti-EM48
Autoradiograph

b Rate of HA-Q79 degradation
(change in [35 S]HA-Q79 in 23 h)
* $P < 0.05$
 $n = 3$
- +
Congo red

c Peptidylglutamyl hydrolysing activity
z-LLE-AMC
(arbitrary units $\times 10^{-3} \text{ min}^{-1}$)
* $P < 0.05$
 $n = 3$
PBS Congo red zVAD
Congo red

d Chymotrypsin-like activity
suc-LLVY-AMC
(arbitrary units $\times 10^{-3} \text{ min}^{-1}$)
* $P < 0.05$
 $n = 3$
PBS Congo red zVAD
Congo red

e Sucrose gradient
5% 40%
 M_r (K) 1 3 6 9 12 15 18
WB: HA 20S
97 46 21
HA 20S
97 46 21
HA 20S
97 46 21
Congo red

f Congo red - +
IgG
WB: IP: HA 20S
20S HA HA 20S
20S HA 20S

NATURE | VOL 421 | 23 JANUARY 2003 | www.nature.com/nature

a

WT HD-1 Tg R62

% of weight change (weeks 9–11)

Vehicle Vehicle Congo red Vehicle Congo red Congo red

* $P < 0.05$

b

HD-1 Tg R62

Vehicle (PBS) i.p. (Congo red)

c

Latency to fall at 10 r.p.m. (s)

Postnatal days

63 70 77 84

* $P < 0.05$
** $P < 0.01$

— i.c.v. and i.p. (Congo red)
- - i.p. (Congo red)
- - i.c.v. (Congo red)
— i.p. (PBS)

d

% of survival

Postnatal days

42 56 70 84 98 112

Vehicle $\bar{X} = 91 \pm 6$
Congo red $\bar{X} = 106 \pm 4$

e 9 wks AB: anti-EM48

f i.c.v.-PBS

g Basal ganglia Hippocampus i.p.-PBS

h i.c.v. Congo red

i i.p. Congo red

j

Mutant full-length PolyQ-containing protein

Cleavage

Congo red

PolyQ

PolyQ oligomers

Degradation

Protein recruitment

Accumulation

Transcription inhibition

Caspase activation

ATP depletion

Protein synthesis inhibition

Proteasome inhibition

Protein aggregates

Cell death

Polishing Group

were subjected to a 5–40% sucrose gradient, soluble HA–Q79 was detected at the top of the gradient and as a ~90K protein on SDS–PAGE. In samples treated with Congo red, a significant portion of soluble HA–Q79 migrated with a high M_r complex at the bottom half of the gradient and as a ~46K protein on SDS–PAGE (Fig. 5f). Notably, the 20S subunit of proteasome was detected in the same fraction and was coimmunoprecipitated with HA–Q79 in samples treated with Congo red (Fig. 5e). These results suggest that Congo red treatment shifts the equilibrium of soluble expanded polyglutamine from the higher oligomeric form to a lower M_r form that permits it to interact with a high M_r complex containing the 20S proteasome subunit.

To examine directly the role of polyglutamine aggregates on characteristic features of disease progression in expanded polyglutamine disorders, we used the R62 mouse model of Huntington's disease, in which a truncated form of huntingtin with 139 CAG repeats is expressed⁷. Because Congo red can disrupt preformed oligomers, we concentrated on its impact on disease progression after the formation of polyglutamine aggregates and the onset of symptoms. We infused Congo red either at a dose of 1 mg per 30 g (body weight) every 48 h intraperitoneally (i.p.) or through a 28-d intracerebroventricular cannula (i.c.v.; 6 μ g every 24 h) starting at postnatal week 9. Infusion of Congo red either i.p. or i.c.v. resulted in no gross morphological abnormalities or induction of any obvious symptoms in normal wild-type mice or in pre- or post-symptomatic R62 mice (data not shown). A common feature of disorders caused by expanded polyglutamine repeats, including Huntington's disease, is severe weight loss as a part of the systemic pathology (Fig. 6a, b). The severe weight loss observed in R62 transgenic mice expressing the expanded polyglutamine repeats was significantly ameliorated after Congo red treatment by either i.c.v. or i.p. delivery (Fig. 6a).

R62 mice also develop severe diabetes owing to the presence of polyglutamine aggregates in the pancreas²⁵, a feature mimicking the elevated diabetes rate in individuals affected with Huntington's disease²⁶. R62 transgenic mice treated with Congo red had significantly reduced blood glucose concentrations (178 ± 46 mg dl⁻¹) as compared with PBS-treated R62 transgenic mice (398 ± 98 mg dl⁻¹) and had similar concentrations to those of wild-type mice treated with either PBS (152 ± 27 mg dl⁻¹) or Congo red (216 ± 60 mg dl⁻¹) after 6 h of fasting. Thus, the treatment of Congo red is effective against both peripheral symptoms of R62 mice.

Initial abnormal neurological signs of R62 mice include dyskinesia of the hindlimbs when mice are suspended by the tail and irregular gait⁷. Treatment with Congo red significantly inhibited the dyskinesia of hindlimbs (Fig. 6b) and preserved normal gait and stride length. Whereas the average stride length of vehicle-treated R62 transgenic mice was decreased by 46% to that of wild type, R62 mice treated with Congo red showed only a 17% reduction in stride length (see Supplementary Fig. 5). The effect of Congo red on motor performance in R62 mice was also assessed by rotarod. R62 mice treated with Congo red preserved their motor function, whereas PBS-treated R62 mice continued to deteriorate (Fig. 6c). The life span of R62 mice was also significantly prolonged by Congo red treatment, resulting in a mean survival length of 106 d as compared with 91 d in vehicle-treated mice (Fig. 6d).

To verify that Congo red could disrupt and inhibit the formation of expanded polyglutamine oligomers *in vivo*, we analysed brain samples from control and Congo-red-treated R62 mice by immunostaining with EM48 antibodies against polyglutamine repeats^{27,28}. The basal ganglia of 9-week-old R62 mice showed extensive EM48-positive polyglutamine aggregates as described previously (ref. 7 and Fig. 6e). Extensive clearance of expanded polyglutamine repeats was observed in the basal ganglia of 12.5-week-old mice treated with Congo red but not with vehicle (Fig. 6, compare panels f and g with h and i). A moderate reduction of

EM48-positive polyglutamine aggregates was also observed in the hippocampus (data not shown). Thus, *in vivo* Congo red can both disrupt preformed polyglutamine aggregates and inhibit the nucleation of new polyglutamine aggregates.

We conclude that disrupting polyglutamine oligomerization directly inhibits the ability of expanded polyglutamine to induce several cytotoxic events, such as activation of caspases and ATP depletion, providing a mechanism of protection independent of cellular metabolism. At the same time, inhibition of polyglutamine oligomerization facilitates the degradation of expanded polyglutamine by increasing its accessibility to proteasome and thereby indirectly reducing the accumulation of toxic products and preserving the normal cellular metabolic activity (Fig. 6j). Although our studies do not address the role of mutant full-length proteins containing expanded polyglutamine repeats, they indicate that inhibition of polyglutamine oligomerization may provide a viable therapeutic approach for Huntington's disease and other related polyglutamine expansion disorders in which toxic expanded polyglutamine fragments are generated. □

Methods

Assays for cellular metabolism

Proteasome activity was determined by using 10 μ M suc-LLVY-AMC or Z-LLE-AMC (Alexis) with or without MG132. Fluorescence for AMC (excitation 380 nm, emission 460 nm) in MG132-treated samples was subtracted as background. The luciferase activity was determined in cells expressing HA–Q79– or pcDNA6–luciferase 48 h after transfection using the Steady-Glo substrate and Wallach plate reader (Packard). We determined total protein synthesis from cells expressing HA–Q79–GFP or GFP, or nontransfected primary neurons after a 1-h incubation with [³⁵S]methionine and 10 μ M MG132 6 h after transfection or after 2 or 7 d after treatment. Proteins were precipitated with 20% trichloroacetic acid and the amount of ³⁵S was determined using a Beckman LS6000 scintillation counter.

Fluorometric determination of caspase activity

Caspase-8-like or caspase-3-like activity was determined by using the AFC-conjugated fluorogenic substrates IETD and DEVD, respectively (Biomol) or *in situ* by the Caspatag kit (Intergen).

Antibodies

We used the following antibodies: EM48 (ref. 27), antibodies against HSP40 and HSP70 (Stressgen) and a rat monoclonal antibody against caspase-8 (ref. 14). Immunoreactivity was determined with the appropriate horseradish-peroxidase-labelled secondary antibody by ECL (Amersham).

Chemical absorption assays

We scanned each compound solution between 200 nm and 800 nm using a DU640 spectrophotometer (Beckman) to determine its optimal 'signature wavelength'. The absorbance of Congo red at 460 nm after incubation with glutathione beads alone was used as the 100% value. Purified recombinant GST–Q19, GST–Q62 and GST–Q81 (ref. 29) were diluted to a final concentration of 5 μ g ml⁻¹. Compounds tested included Congo red, minocycline, dextran, iodoacetamide, rolitetracycline (Sigma) chrysamine G (AG Scientific) or vehicle alone (0.2% dimethyl sulphoxide (DMSO) in PBS).

Cell-free aggregation and filter assay

HeLa cell lysates were passed through a needle to ensure the release of polyglutamine inclusions from the nuclei. The supernatants were centrifuged twice at 4,000 r.p.m. for 30 s over a 25% sucrose cushion to remove unbroken nuclei. Polyglutamine aggregates were aliquoted into 384-well plates and treated with Congo red, minocycline or rolitetracycline at a final concentration of 0–100 μ M. Filter assays were done as described¹⁰. The HA–Q79–GFP aggregates and recombinant GST–Q19 and GST–Q81 were passed through a 0.2- μ m filter and the presence of HA-tagged Q79 or polyglutamine was detected with antibodies against HA or with EM48 antibodies, respectively.

Immunoprecipitation and protein turnover rates

HeLa cells transfected with a HA–Q79 construct were treated with 100 μ M Congo red or PBS 6 h after transfection. Twenty-three hours after transfection, cells expressing HA–Q79 were metabolically labelled in methionine-free DMEM medium containing 50 μ Ci of [³⁵S]methionine (Gibco) in the presence of PBS or 100 μ M Congo red for 1 h. We washed cells to remove unincorporated label and chased them for 1 or 24 h before collection. After centrifugation at 2,000g for 3 min, the cells were lysed in PBS containing Triton X-100 and protease inhibitors. Sepharose bound to HA antibodies (Babco) was used to immunoprecipitate HA–Q79 from lysates. We processed immunoprecipitates for autoradiography and western blotting with EM48 antibodies. To determine the rates of total cellular protein degradation, 2×10^4 cells were seeded in 12-well dishes and treated with Congo red (100 μ M), zVAD.fmk (100 μ M) or MG132 (10 μ M) and labelled with 5 μ Ci ml⁻¹ [³H]tyrosine for 1 h and the amounts of [³H]tyrosine determined using the LS6000 scintillation counter (Beckman).

Mice surgery and treatment

We obtained 6-week-old R62 transgenic mice and CBA $\times$ C57Bl/6 F₁ wild-type littermates from Jackson laboratories. The genotype was confirmed by polymerase chain reaction as described⁷. Mice were treated after 1 week of daily rotarod training. Mice were anaesthetized by intraperitoneal injection of chloral hydrate and osmotic pumps (0.25 μ l h⁻¹ for 28 d), and the cannula were implanted intracerebroventricularly (Alzet) using predetermined coordinates (anterior/posterior, -0.5 mm, 1 mm lateral to the bregma). Congo red was diluted at 1 mg ml⁻¹ in calcium- and magnesium-free PBS plus 0.2% DMSO. The vehicle solution contained 0.2% DMSO in calcium- and magnesium-free PBS. For intraperitoneal injections, 0.5 ml of 1 mg ml⁻¹ Congo red in PBS plus 0.2% DMSO was injected every 48 h. The protocol was approved by the Harvard Medical School Standing Committee on Animals.

Behavioural tests, rotarod performance

Two days after their first treatment, we tested the mice for motor performance and coordination by using a rotarod (Columbus Instruments) at 10 r.p.m. for a maximum of 210 s and by the 'ink' test³⁰. The mice were weighed once a week. Two rotarod trials were carried out three times a week in a blind manner.

Tissue preparation and histology

Mice were anaesthetized with isoflurane and perfused intracardially with 4% paraformaldehyde in PBS. The brain was removed and washed several times in PBS before being incubated overnight in PBS containing 30% sucrose and embedded in OCT (optimal cutting temperature) (Sigma) for sectioning. We stained the sections with EM48 antibodies at a dilution of 1:1,000 and detected immunoreactivity with the ABC kit according to the manufacturer's instructions (Vector).

Received 14 August; accepted 6 November 2002; doi:10.1038/nature01301.

1. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
2. Wellington, C. L. *et al.* Inhibiting caspase cleavage of huntingtin reduces toxicity and aggregate formation in neuronal and nonneuronal cells. *J. Biol. Chem.* **275**, 19831–19838 (2000).
3. Kim, Y. J. *et al.* Caspase 3-cleaved N-terminal fragments of wild-type and mutant huntingtin are present in normal and Huntington's disease brains, associate with membranes, and undergo calpain-dependent proteolysis. *Proc. Natl Acad. Sci. USA* **98**, 12784–12789 (2001).
4. Lunkes, A. *et al.* Proteases acting on mutant huntingtin generate cleaved products that differentially build up cytoplasmic and nuclear inclusions. *Mol. Cell* **10**, 259–269 (2002).
5. Wellington, C. L. *et al.* Caspase cleavage of mutant huntingtin precedes neurodegeneration in Huntington's disease. *J. Neurosci.* **22**, 7862–7872 (2002).
6. Perutz, M. Polar zippers: their role in human disease. *Protein Sci.* **3**, 1629–1637 (1994).
7. Mangiarini, L. *et al.* Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* **87**, 493–506 (1996).
8. Klunk, W. E., Pettigrew, J. W. & Abraham, D. J. Quantitative evaluation of Congo red binding to amyloid-like proteins with a β -pleated sheet conformation. *J. Histochem. Cytochem.* **37**, 1273–1281 (1989).
9. Carter, D. B. & Chou, K. C. A model for structure-dependent binding of Congo red to Alzheimer β -amyloid fibrils. *Neurobiol. Aging* **19**, 37–40 (1998).
10. Heiser, V. *et al.* Inhibition of huntingtin fibrillogenesis by specific antibodies and small molecules: implications for Huntington's disease therapy. *Proc. Natl Acad. Sci. USA* **97**, 6739–6744 (2000).
11. Caughey, B. & Raymond, G. J. Sulfated polyanion inhibition of scrapie-associated PrP accumulation in cultured cells. *J. Virol.* **67**, 643–650 (1993).
12. Howlett, D. R., George, A. R., Owen, D. E., Ward, R. V. & Markwell, R. E. Common structural features determine the effectiveness of carvedilol, daunomycin and retitetracycline as inhibitors of Alzheimer β -amyloid fibril formation. *Biochem. J.* **343**, 419–423 (1999).
13. Klunk, W. E., Debnath, M. L. & Pettigrew, J. W. Chrysamine-G binding to Alzheimer and control brain: autopsy study of a new amyloid probe. *Neurobiol. Aging* **16**, 541–548 (1995).
14. Sanchez, I. *et al.* Caspase-8 is required for cell death induced by expanded polyglutamine repeats. *Neuron* **22**, 623–633 (1999).
15. Beal, M. F., Hyman, B. T. & Koroshetz, W. Do defects in mitochondrial energy metabolism underlie the pathology of neurodegenerative diseases? *Trends Neurosci.* **16**, 125–131 (1993).
16. Deckwerth, T. L. & Johnson, E. M. Jr Temporal analysis of events associated with programmed cell death (apoptosis) of sympathetic neurons deprived of nerve growth factor. *J. Cell Biol.* **123**, 1207–1222 (1993).
17. Jaattela, M., Wissing, D., Kokholm, K., Kallunki, T. & Egeblad, M. Hsp70 exerts its anti-apoptotic function downstream of caspase-3-like proteases. *EMBO J.* **17**, 6124–6134 (1998).
18. Wyttenbach, A. *et al.* Effects of heat shock, heat shock protein 40 (HDJ-2), and proteasome inhibition on protein aggregation in cellular models of Huntington's disease. *Proc. Natl Acad. Sci. USA* **97**, 2898–2903 (2000).
19. Chai, Y., Koppenhafer, S. L., Bonini, N. M. & Paulson, H. L. Analysis of the role of heat shock protein (Hsp) molecular chaperones in polyglutamine disease. *J. Neurosci.* **19**, 10338–10347 (1999).
20. Kazemi-Esfarjani, P. & Benzer, S. Genetic suppression of polyglutamine toxicity in *Drosophila*. *Science* **287**, 1837–1840 (2000).
21. Hazeki, N., Takamoto, T., Goto, J. & Kanazawa, I. Formic acid dissolves aggregates of an N-terminal huntingtin fragment containing an expanded polyglutamine tract: applying to quantification of protein components of the aggregates. *Biochem. Biophys. Res. Commun.* **277**, 386–393 (2000).
22. Moulder, K. L., Onodera, O., Burke, J. R., Strittmatter, W. J. & Johnson, E. M. Jr Generation of neuronal intranuclear inclusions by polyglutamine-GFP: analysis of inclusion clearance and toxicity as a function of polyglutamine length. *J. Neurosci.* **19**, 705–715 (1999).
23. Cummings, C. J. *et al.* Chaperone suppression of aggregation and altered subcellular proteasome localization imply protein misfolding in SCA1. *Nature Genet.* **19**, 148–154 (1998).
24. Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* **292**, 1552–1555 (2001).
25. Hurlbert, M. S. *et al.* Mice transgenic for an expanded CAG repeat in the Huntington's disease gene develop diabetes. *Diabetes* **48**, 649–651 (1999).

26. Farrer, L. A. Diabetes mellitus in Huntington disease. *Clin. Genet.* **27**, 62–67 (1985).
27. Li, H., Li, S. H., Johnston, H., Shelbourne, P. F. & Li, X. J. Amino-terminal fragments of mutant huntingtin show selective accumulation in striatal neurons and synaptic toxicity. *Nature Genet.* **25**, 385–389 (2000).
28. DiFiglia, M. *et al.* Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science* **277**, 1990–1993 (1997).
29. Onodera, O. *et al.* Oligomerization of expanded-polyglutamine domain fluorescent fusion proteins in cultured mammalian cells. *Biochem. Biophys. Res. Commun.* **238**, 599–605 (1997).
30. Carter, R. J. *et al.* Characterization of progressive motor deficits in mice transgenic for the human Huntington's disease mutation. *J. Neurosci.* **19**, 3248–3257 (1999).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank E. Signer, S. Lokey, M. Kirschner, N. Ayad, M. Kobori, O. Gozani, L. Yoo, R. Sanchez-Olea, A. Degterev and R. King for comments; T. Mitchison for advice; M. Takeuchi for technical assistance; and X.-J. Li, A. Kazantsev and C. Cepko for reagents. This work was supported in part by grants from the Hereditary Disease Foundation and the NIH (to J.Y.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.Y. (e-mail: Junying_Yuan@hms.harvard.edu).

A role for *Drosophila* LKB1 in anterior–posterior axis formation and epithelial polarity

Sophie G. Martin & Daniel St Johnston

The Wellcome Trust/Cancer Research UK Institute and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK

The PAR-4 and PAR-1 kinases are necessary for the formation of the anterior–posterior (A–P) axis in *Caenorhabditis elegans*^{1–3}. PAR-1 is also required for A–P axis determination in *Drosophila*^{4,5}. Here we show that the *Drosophila* *par-4* homologue, *lkb1*, is required for the early A–P polarity of the oocyte, and for the repolarization of the oocyte cytoskeleton that defines the embryonic A–P axis. LKB1 is phosphorylated by PAR-1 *in vitro*, and overexpression of LKB1 partially rescues the *par-1* phenotype. These two kinases therefore function in a conserved pathway for axis formation in flies and worms. *lkb1* mutant clones also disrupt apical–basal epithelial polarity, suggesting a general role in cell polarization. The human homologue, LKB1, is mutated in Peutz–Jeghers syndrome^{6,7} and is regulated by prenylation and by phosphorylation by protein kinase A^{8,9}. We show that protein kinase A phosphorylates *Drosophila* LKB1 on a conserved site that is important for its activity. Thus, *Drosophila* and human LKB1 may be functional homologues, suggesting that loss of cell polarity may contribute to tumour formation in individuals with Peutz–Jeghers syndrome.

The A–P axis of *Drosophila* is specified during oogenesis when a signal from the posterior follicle cells induces the formation of a polarized oocyte microtubule cytoskeleton, in which most minus ends are nucleated from the anterior cortex, with the plus ends extending towards the posterior pole¹⁰. These polarized microtubules direct both the localization of *bicoid* messenger RNA to the anterior of the oocyte to specify where the head and thorax will develop, and the transport of *oskar* mRNA to the posterior, where it induces the formation of polar granules that contain the abdominal and germline determinants¹⁰.

To identify other genes required for formation of the A–P axis, we carried out a genetic screen in germline clones for mutants that disrupt the localization of Staufen tagged with green fluorescent protein (GFP), which colocalizes with *bicoid* and *oskar* mRNAs¹¹. Mutants in one lethal complementation group of two alleles abolish the posterior localization of both Staufen and *oskar* mRNA in 80–90% of oocytes ($n = 207$; Fig. 1a, b, i, j). In over 40% of mutant oocytes, both *bicoid* and *K10* mRNAs are detected around the whole oocyte cortex, rather than only at the anterior pole ($n = 309$; Fig. 1c, d, and data not shown). In contrast to wild-type oocytes, which have an anterior-to-posterior gradient of microtubules, mutant oocytes have a high density of microtubules all around the cortex, with the lowest concentration in the centre (Fig. 1e, f). Kinesin- β -galactosidase, a microtubule plus-end marker¹², also fails to concentrate at the posterior of the oocyte and accumulates instead in the centre (98%, $n = 80$; Fig. 1g–j). Thus, these mutants disrupt *bicoid* and *oskar* mRNA localization by preventing the A–P polarization of the microtubules.

Using a mapping strategy based on single-nucleotide polymorphisms, we located this complementation group to a 46-kilobase (kb) fragment in 87F (ref. 13). The sequencing of candidate genes showed that both alleles contain lesions in CG9374, which are predicted to be null mutations (Fig. 2a). A genomic rescue construct rescues both lethality and localization of *oskar* mRNA to the posterior of the oocyte, confirming that these phenotypes are caused by mutations in this gene. CG9374 encodes a serine/threonine kinase with homology to *C. elegans* PAR-4 (ref. 2), a kinase involved in the localization of P granules to the posterior of

the one-cell zygote¹, and to the human tumour-suppressor LKB1 (refs 6, 7, and Fig. 2a). Given the stronger homology to the latter, we named the *Drosophila* gene *lkb1*.

Hypomorphic mutants in *Drosophila par-1* show defects in the polarization of the oocyte microtubule cytoskeleton and in the localization of *bicoid* and *oskar* mRNA that are very similar to the defects of *lkb1* mutants^{4,5,14}. The PAR-1 kinase is also required much earlier in oogenesis for the determination of the oocyte. The oocyte is selected from a cyst of 16 germline cells in the germarium and forms a microtubule-organizing centre, which directs the microtubule-dependent localization of oocyte-specific factors, such as ORB, to this cell¹⁵. The microtubule-organizing centre then moves from the anterior to the posterior of the oocyte in region 3, the most posterior region of the germarium, and oocyte-specific factors accumulate posteriorly¹⁶. This anterior-to-posterior switch is disrupted in *par-1* null mutants, and the oocyte exits meiosis and becomes a nurse cell^{16,17}.

To study the role of *lkb1* in this process, we induced mutant germline clones marked by the loss of GFP. ORB still accumulates in the presumptive oocyte in *lkb1* cysts but often fails to move to the posterior in region 3; instead, it disperses throughout the cyst as the oocyte exits meiosis and adopts the nurse cell fate (Fig. 1k, l). Thus, *lkb1* and *par-1* share very similar phenotypes in both oocyte determination and polarization, suggesting that they function together. Because the penetrance of the early *lkb1* phenotype increases with age (67% ($n = 68$) of mutant clones in 5-day-old females, rising to 85% ($n = 73$) after 12 d), wild-type LKB1 activity seems to perdure for several days after the clones are induced, and

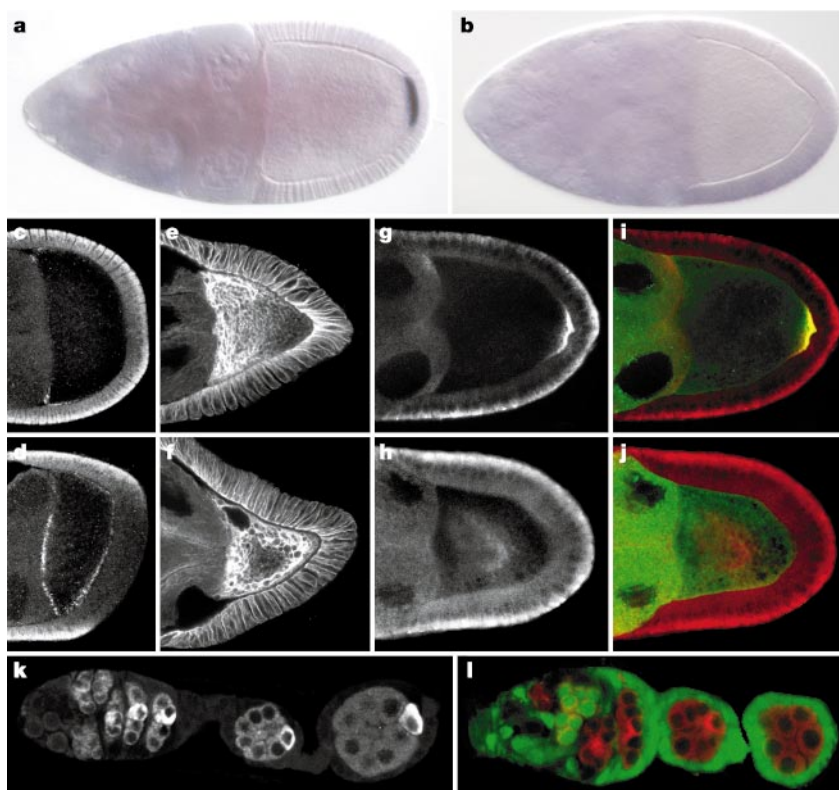


Figure 1 *lkb1* mutants affect oocyte polarity. **a, c, e, g, i, k**, Wild type. **b, d, f, h, j, l**, Mutant germline clones. Both alleles have an identical phenotype and therefore only one allele is shown (4A4-2: **b, d, f, i**; 4B1-11: **h, j**). **a, b**, *In situ* hybridization for *oskar* mRNA. Staufen and *oskar* mRNA accumulate in the oocyte but fail to localize to the posterior in 80–90% of the cases. The residual posterior localization in 10–20% of the oocytes is probably due to perdurance of the protein, as shown for the oocyte determination phenotype. The resulting embryonic phenotype could not be scored because embryos

arrest before cellularization. **c, d**, Fluorescent *in situ* hybridization for *bicoid* mRNA. **e, f**, α -Tubulin staining. **g, h**, Kinesin- β -gal localization, detected with an antibody against β -gal. **i, j**, Overlay of kinesin- β -gal (red) and GFP-Staufen (green). Colocalization is shown in yellow. **k**, ORB staining in wild type. **l**, ORB staining (red) in homozygous 4A4-2 clones, marked by the absence of GFP (green). ORB accumulates in the oocyte but fails to relocate to the posterior. Anterior is to the left, posterior is to the right in all figures.

this presumably accounts for the escapers that show the later oocyte polarity phenotype.

As PAR-1 and LKB1 seem to act in a common pathway, we tested whether either kinase could phosphorylate the other. Although recombinant PAR-1 is not a substrate for LKB1, immunoprecipitated GFP-PAR-1 phosphorylates recombinant LKB1 *in vitro* (Fig. 3a). We found that the phosphorylation site or sites are located in the amino-terminal half of the protein, although we have been unable to map them precisely. We also observed a strong genetic interaction between the two genes. The *oskar* mRNA localization defects in hypomorphic combinations of *par-1* alleles cause loss of abdominal segments in the embryo, which can be strongly enhanced by removing one copy of *lkb1* (Supplementary Fig. 1a). In addition, overexpression of GFP-LKB1 partially rescues the A-P polarity phenotype of *par-1*⁶³²³/*par-1*^{W3}, the strongest allelic combination that produces late-stage oocytes. Whereas Staufer localizes to the posterior normally in only 12% of these oocytes, 80% have wild-type amounts of Staufer at the posterior when LKB1 is overexpressed (Supplementary Fig. 1b). This phenotype is not rescued by a kinase-dead form of LKB1 (Supplementary Fig. 2), indicating that this suppression requires kinase activity. By contrast, overexpression of PAR-1 does not rescue the phenotype of *lkb1* mutant germline clones (data not shown). These results are consistent with a model in which LKB1 is a direct target of PAR-1 regulation *in vivo* and functions as a downstream effector in the polarization of the oocyte microtubule cytoskeleton.

A GFP-LKB1 fusion construct under the control of the endogenous promoter rescues both the lethality and oogenesis phenotypes of *lkb1* mutants and is expressed in very low amounts in both the germline and somatic follicle cells of the ovary (Fig. 2c). The highest expression was observed in the germarium, where GFP-LKB1 colocalizes with PAR-1 on the fusome^{16,17}, a branched membranous organelle that connects the germ cells in a cyst (ref. 15 and Fig. 2d). This localization presumably reflects their common function in early oocyte polarity and determination, because the cell that inherits most fusome is selected to become the oocyte¹⁸. During the rest of oogenesis, GFP-LKB1 shows a uniform cortical localization in both the germline and the follicle cells (Fig. 2c, f, g). It is enriched in the oocyte from stage 7, when the A-P axis is polarized, and colocalizes with cortical actin, but not with pole plasm components (Fig. 2e).

The centrally located kinase domain of LKB1 has 44% and 66% amino-acid identity with PAR-4 and human LKB1, respectively, but the amino- and carboxy-terminal regions are not well conserved (Fig. 2a). *Drosophila* and human LKB1 share a C-terminal prenylation motif (Fig. 2b), however, and the latter is prenylated *in vivo*, which is necessary for its localization to the plasma membrane^{8,9}. Mutation of the prenyl-acceptor cysteine to alanine (C537A) in *Drosophila* GFP-LKB1 results in a cytoplasmic accumulation of the protein, showing that *Drosophila* LKB1 is targeted to the cell cortex through prenylation (100%, *n* = 71; Fig. 2g, h). This cortical localization is essential for LKB1 function, because UAS-GFP-LKB1^{C537A} does not rescue the *lkb1* phenotype when expressed in near-endogenous amounts (Fig. 3c). This construct can, however, partially rescue the posterior localization of Staufer when it is expressed in tenfold higher amounts (Fig. 3d). In agreement with this observation, the LKB1^{C537A} mutant shows a weak residual membrane localization, which may be sufficient to mediate LKB1 function when overexpressed. Thus, cortical localization is important for LKB1 activity, consistent with its role in organizing microtubules along the lateral and posterior cortex.

Drosophila LKB1 also has a conserved RKLS consensus phosphorylation site near its C terminus (Fig. 2b). In mammalian LKB1, this site is phosphorylated *in vitro* and *in vivo* by protein kinase A (PKA) and is required for its ability to suppress cell growth in culture^{8,9}. Like its vertebrate counterparts, *Drosophila* LKB1 is phosphorylated in a PKA-dependent manner. Coexpression of

wild-type LKB1 and PKA in S2 cells induces a phosphatase-sensitive mobility shift of LKB1 on western blots, which is abolished when serine 535 is mutated to alanine (Fig. 3b). To assay the significance of phosphorylation of this conserved site, we generated transgenes in which the serine was mutated to either alanine (S535A) to prevent phosphorylation, or to glutamic acid (S535E) to mimic the presence

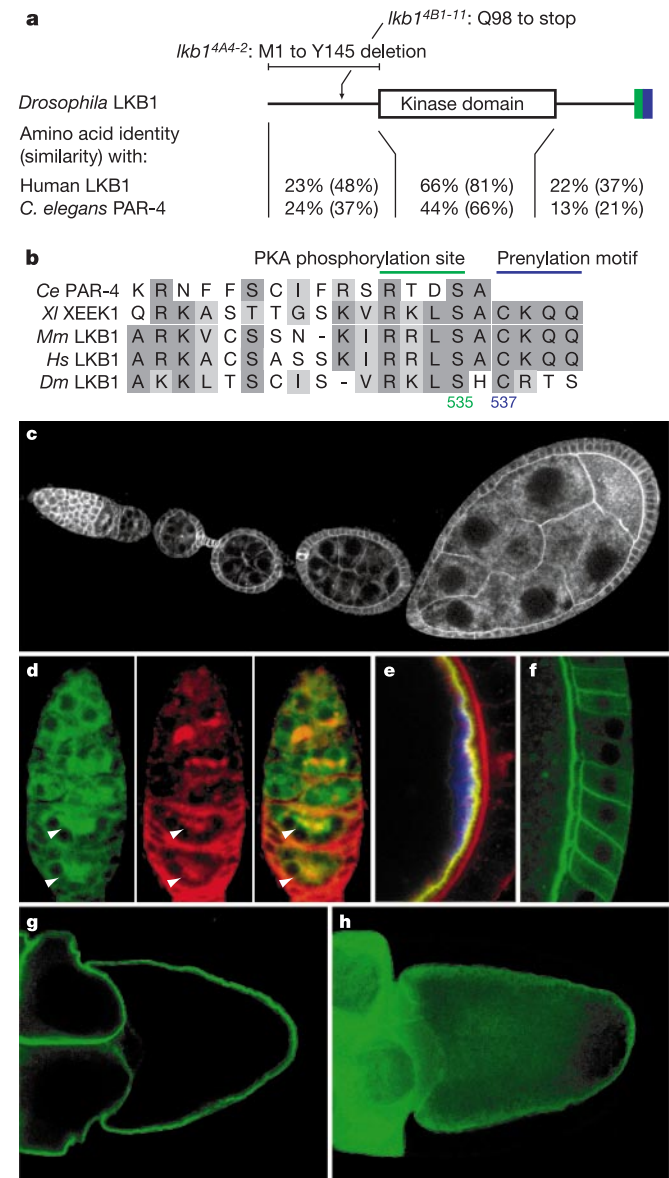


Figure 2 Conservation and localization of LKB1. **a**, *lkb1*^{4A4-2} contains a 589-bp deletion (associated with a 12-bp insertion) removing 150 bp of the 5' untranslated region, the start codon and the beginning of the ORF. *lkb1*^{4B1-11} contains a nonsense mutation at amino acid 98. The conserved kinase, phosphorylation (green) and prenylation (blue) domains are indicated. **b**, Alignment of the C termini of LKB1 homologues, showing the phosphorylation (S535) and prenylation (C537) motifs conserved between *Drosophila* and human LKB1. **c**, Expression of the GFP-LKB1 genomic transgene showing endogenous expression of *lkb1*. Because expression is very low, this image was taken at 100% laser power and the diffuse cytoplasmic staining is background. **d**, GFP-LKB1 (green) and PAR-1 (red) colocalize on the fusome in regions 2b and 3 of the germarium (arrowheads). **e**, GFP-LKB1 (green) colocalizes with actin (red) but not with Staufer (blue) in stage-10 oocytes. UAS-GFP-LKB1 was driven by *matα4*-GAL4-VP16. **f**, GFP-LKB1 localizes cortically in follicle cells. UAS-GFP-LKB1 was driven by *da*-GAL4. **g**, Expression of wild-type GFP-LKB1. **h**, Expression of GFP-LKB1^{C537A}, which cannot be prenylated. In **f** and **g**, both transgenes were expressed at similar levels, driven by *matα4*-GAL4-VP16.

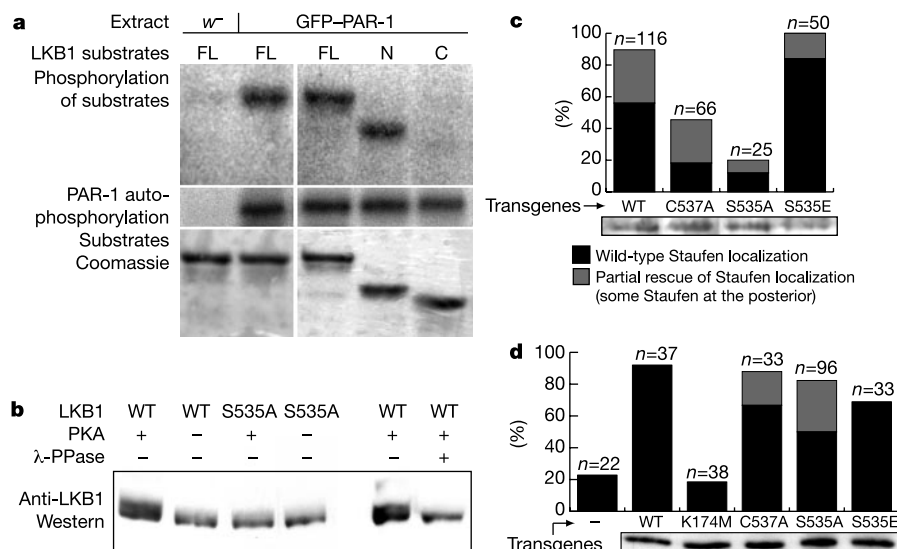


Figure 3 Analysis of LKB1 regulation. **a**, Immunoprecipitated GFP-PAR-1 phosphorylates the N-terminal half of recombinant LKB1 *in vitro*. As a control, immunoprecipitation was also carried out with a w^- extract. PAR-1 autophosphorylation shows equal levels of PAR-1 activity. FL, LKB1 full length; C, C terminus; N, N terminus. **b**, Western blot of extracts of S2 cells transfected with LKB1 and PKA as indicated and detected with antibodies against LKB1. Coexpression of LKB1 and PKA induces a mobility shift in LKB1, which is dependent on S535 and abolished by phosphatase treatment. **c**, Rescue of the posterior localization of Staufen in *lkb1^{4A4-2/lkb1^{4B1-11}}* flies by expression of UAS-GFP-LKB1. Transgenes were expressed with *arm-GAL4*, and *lkb1^{4A4-2/lkb1^{4B1-11}}* flies were recovered expressing low levels of *lkb1* in the germ line. GFP-LKB1^{C537A} and

GFP-LKB1^{S535A} transgenes show impaired function; by contrast, GFP-LKB1^{S535E} rescues more efficiently than wild-type. Western blot with antibodies against GFP shows that all transgenes are expressed in similar quantities. For S535A, only egg chambers with wild-type morphology were scored (see Fig. 4). **d**, Overexpression of wild-type and mutant transgenes with *mat α 4-GAL4-VP16* in *lkb1^{4A4-2}* germline clones. The kinase-dead mutant GFP-LKB1^{K174M} (see Supplementary Fig. 2) is unable to rescue, whereas the phosphorylation (GFP-LKB1^{S535A}) and prenylation (GFP-LKB1^{C537A}) mutants rescue LKB1 function when overexpressed. Western blot with antibodies against GFP shows that all transgenes are expressed in similar amounts, about tenfold more than endogenous protein.

of a charged phosphate group. Expression of the transgenes with *arm-GAL4* allows the recovery of *lkb1* mutant flies expressing low amounts of either wild-type or mutant proteins in the germ line¹⁹. GFP-LKB1^{S535A} does not rescue the localization of Staufen to the posterior of the oocyte, whereas GFP-LKB1^{S535E} rescues even more efficiently than the wild-type control, strongly suggesting that LKB1 is positively regulated by phosphorylation at this site (Fig. 3c). Because S535 is phosphorylated by PKA *in vivo* and PKA is required

in the germ line to polarize the oocyte²⁰, we speculate that PKA regulates *Drosophila* LKB1 in the germ line. Additional signals must regulate LKB1, however, because the lack of phosphorylation on S535 does not abolish LKB1 activity completely. Tenfold overexpression of S535A in *lkb1* germline clones partially rescues the localization of Staufen to the posterior of the oocyte, albeit less efficiently than the wild-type transgene, whereas a kinase-dead version (K174M) shows no rescuing activity on overexpression

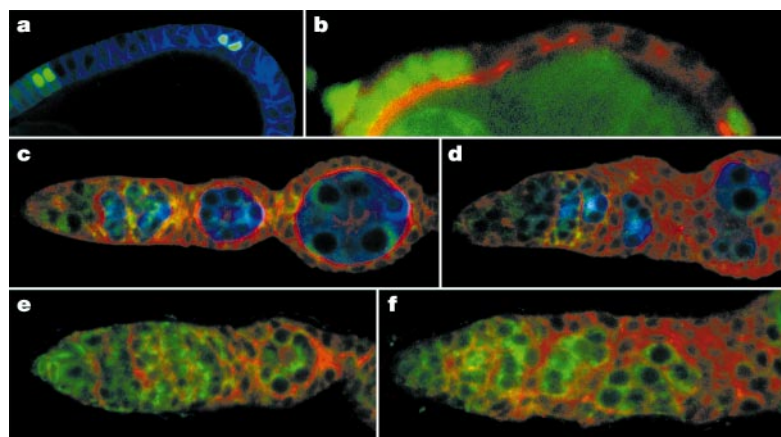


Figure 4 *lkb1* is required for epithelial polarity. **a, b**, *lkb1^{4A4-2}* follicle cell clones marked by the absence of GFP (green) and stained for α -spectrin (**a**, blue) and aPKC (**b**, red). **c, d**, GFP-LKB1^{S535E} (**c**, green) but not GFP-LKB1^{S535A} (**d**, green) rescues the apical localization of aPKC (red) in the follicular epithelium. ORB (blue) stains the germ line. **e, f**, Similarly, wild-type GFP-LKB1 (**e**, green) but not GFP-LKB1^{S535A} (**f**, green) rescues the apical localization of Armadillo (red), a component of adherens junctions. Most ovarioles expressing GFP-LKB1^{S535A} show aberrant morphology of the germarium (**d, f**),

a phenotype that is also observed with large somatic *lkb1* clones (not shown). In **c-f**, transgenes were driven by *arm-GAL4* in *lkb1^{4A4-2/lkb1^{4B1-11}}* flies. LKB1 has been implicated in apoptosis and cell-cycle control²³, but the polarity phenotype observed was not a consequence of cell death or overproliferation because TUNEL assay and phosphohistone H3 staining failed to detect more apoptotic or mitotic cells in mutant clones than in their wild-type neighbours (not shown). Thus, *lkb1* seems to have a specific role in the polarity of epithelial cells.

(Fig. 3d and Supplementary Fig. 2). Thus, LKB1 may be regulated by both PAR-1 and PKA, and may function to integrate the two signalling pathways during the polarization of the oocyte.

Follicle cell clones mutant for *lkb1* also show a defect in polarity. In severely affected clones, the follicular monolayer is disorganized, with mutant cells rounding up and sorting out of the epithelium (Fig. 4a). Morphologically wild-type clones show defects in the apical localization of atypical protein kinase C (aPKC) and Armadillo, which become either diffuse or ectopically localized along lateral membranes (Fig. 4b). In less severely affected cells, the apical localization is discontinuous. These phenotypes are penetrant in large stem-cell clones but not in small clones, indicating that LKB1 activity perdures. Expression of the wild-type or S535E transgenes with *arm-GAL4* in *lkb1* mutants rescues these follicle cell phenotypes, whereas expression of S535A rescues lethality but gives rise to a completely disorganized follicular epithelium, in which most cells appear unpolarized (Fig. 4c–f). Thus, LKB1 is required for cell polarity in the germ line and the follicle cells, and is probably regulated by phosphorylation on the conserved C-terminal serine in both processes.

In *Drosophila*, *par-1* and *lkb1*, the homologue of *C. elegans par-4*, show very similar phenotypes^{4,5}. In addition, LKB1 is an *in vitro* substrate for PAR-1 and can suppress the polarity phenotype of *par-1* mutants when overexpressed. These results suggest that LKB1 functions downstream of PAR-1. This conclusion is consistent with genetic data in *C. elegans* that show that *par-4* mutants display only a subset of the *par-1* A–P polarity phenotypes. Notably, mutants in *par-4* and *par-1*, but not in other *par* genes, show a disappearance of P granules in the one-cell zygote¹. Thus, LKB1 and PAR-1 function in a conserved pathway that is required for the polarization of the A–P axis in both worms and flies.

Drosophila LKB1 is closely related to human LKB1, and conserved prenylation and PKA phosphorylation sites are essential for the *in vivo* function of both proteins, indicating that the two are likely to be functional homologues. Mutants in LKB1 cause Peutz–Jeghers syndrome^{6,7}, which is characterized by the formation of intestinal polyps and a high incidence of adeno-carcinomas (tumours of epithelial origin)²¹. In addition, mutations in *lkb1* have been identified in several sporadic epithelial cancers²². The role of LKB1 as a tumour suppressor is not well understood, however, and LKB1 has been proposed to regulate apoptosis, the cell cycle or angiogenesis²³. In addition, LKB1 seems to function in a context-dependent manner that is different from classical tumour-suppressor genes such as *ras* or *p53* (ref. 24). Given that *Drosophila lkb1* is required to polarize the epithelial follicle cells, we propose the alternative model that loss of LKB1 leads to polyp and tumour formation by disrupting epithelial polarity. □

Methods

Fly strains

We used the following strains: *par-1*^{W3}, *par-1*⁶³²³ and *par-1*⁶⁸²¹ (ref. 4). *lkb1*^{4A4-2} and *lkb1*^{4B1-11} were induced by ethyl methyl sulphonate (EMS) on the FRT82B chromosome and identified on the basis of their failure to localize GFP–Staufen to the posterior of the oocyte (S.G.M., V. Leclerc, K. Litère and D.S.J., manuscript in preparation). The FLP–FRT system was used to induce clones²⁵, in combination with either ovo^{D1} (ref. 26) or GFP (a gift from S. Luschnig) for clone selection. GFP–Staufen was expressed under the control of the *matα4*-tubulin promoter. The kinesin-β-galactosidase transgene¹² was mobilized to obtain an insertion on the first chromosome. For rescue of lethality and low expression in the ovaries, we used *armGAL4* (ref. 27). *matα4*-tubulin–GAL4–VP16 was used for overexpression in the germ line. We used da–GAL4 to study the localization of GFP–LKB1 in somatic follicle cells²⁸.

Molecular biology

The genomic rescue construct contains the DNA between the two genes adjacent to *lkb1*. GFP was inserted in-frame upstream of the *lkb1* open reading frame (ORF CG9374) to create the genomic GFP–LKB1 construct. Inducible GFP–LKB1 transgenes were obtained by inserting the 3.6-kb *lkb1* transcription unit, amplified from the genomic GFP construct, into the UASp vector¹⁹. Mutant versions were generated using *in vitro* site-directed mutagenesis based on polymerase chain reaction (PCR) and sequenced. MBP–lkb1 plasmids were obtained by inserting *lkb1* fragments amplified from GH14740 (BDGP; the

ORF is identical to that available under GenBank accession number AY069241) into pMAL (NEB) in-frame downstream of MBP. LKB1 includes amino-acids 1–301 of the CG9374 protein, and LKB1C includes amino acids 272–540. For expression in S2 cells, we cloned PKA in pMT–DEST48 in-frame with the His₆–V5 epitope at the C terminus using Gateway Technology (Invitrogen). LKB1 was also cloned into pMT–DEST48 with the inclusion of a stop codon, so that LKB1 was untagged.

Kinase assay

Ovaries from either *w⁻* or *matα4-GFP-PAR-1^{NIS}* (ref. 16) flies were homogenized in 150 mM NaCl, 50 mM HEPES (pH 7.2), 0.5 mM dithiothreitol and 1 μM microcystin-LR, supplemented with protease inhibitors (Roche). We carried out immunoprecipitations with affinity-purified polyclonal antibodies against GFP bound to magnetic Dynabeads (Dyna). Kinase buffer (50 mM HEPES (pH 7.4), 1 mM EDTA, 5% glycerol, 150 mM NaCl, 10 mM MgCl₂ and 10 mM MnCl₂) and substrate (2 μM) were mixed directly with the beads and the reaction was started by adding [γ-³²P]ATP. After a 20-min incubation at 30 °C, the reaction was stopped by adding Laemmli buffer and analysed by SDS–PAGE. ³²P incorporation was detected by a phosphorimager. Substrates were expressed as maltose-binding protein (MBP) fusion proteins in bacteria, and purified over amylose columns (NEB). The same procedure, without added substrate, was used for autophosphorylation assays, using ovarian extracts from flies carrying *matα4-GAL4-VP16* and UASp–GFP–LKB1 transgenes.

Cell culture and phosphatase treatment

S2 cells were maintained in Schneider's medium supplemented with 5% fetal calf serum (Sigma). We transfected 10⁶ cells per ml with 1 μg of total DNA (0.2 μg of LKB1 and either 0.8 μg of PKA or 0.8 μg of empty pMT–DEST48 vector) with FuGENE 6 (Roche). After 24 h, the culture medium was removed and fresh medium containing 200 μM CuSO₄ was added to induce expression of the metallothionein promoter. We collected the cells 24 h later, and resuspended them in RIPA buffer (150 mM NaCl, 1% Nonidet P-40, 0.5% deoxycholate, 0.1% SDS, 50 mM Tris pH 8.0, 0.4 mM EDTA, 10% glycerol and protease inhibitor cocktail (Roche)). Treatment with λ-phosphatase (NEB) was carried out at 30 °C for 30 min and stopped by the addition of Laemmli buffer.

Immunostaining and *in situ* hybridization

We carried out immunostaining according to standard procedures using the following antibodies: a rabbit polyclonal against Staufén (ref. 11; used at 1:5,000 dilution), a rabbit polyclonal against PAR-1 (ref. 29; 1:1,250), a mouse monoclonal against β-gal (JIE7; 1:400), a mouse monoclonal against ORB (6H4 and 4H8; 1:500), a mouse monoclonal against Armadillo (7A1; 1:200) and a mouse monoclonal against α-spectrin (3A9; 1:1,000) from DSHB; a rabbit polyclonal against aPKC (Santa Cruz; 1:500); and a rabbit polyclonal against phosphohistone H3 (Upstate Biotechnology; 1:1,000). Texas-red-, fluorescein isothiocyanate (FITC)- and Cy5-conjugated secondary antibodies (Molecular Probes) were used at a 1:200 dilution. For microtubule staining, we fixed samples for 10 min with 8% paraformaldehyde in fixation buffer³⁰ and stained them with a FITC-conjugated monoclonal antibody against α-tubulin (Sigma; 1:400). Apoptag Red (Intergen) was used for TdT-mediated dUTP nick end labelling (TUNEL) assays. *In situ* hybridizations were carried out using RNA probes labelled with digoxigenin-UTP (Roche). For fluorescent detection, a secondary polyclonal sheep antibody against DIG (Roche; 1:5,000) and a tertiary Cy3-conjugated donkey antibody against sheep IgG (Jackson Immunoresearch; 1:200) were used. Immunohistochemical detection was done with alkaline-phosphatase-conjugated antibodies against DIG (Roche; 1:5,000). We raised LKB1 antibodies against MBP–LKB1N in rabbits (Eurogentech) and affinity-purified them.

Received 19 June; accepted 6 November 2002; doi:10.1038/nature01296.

- Kemphues, K. J., Priess, J. R., Morton, D. G. & Cheng, N. S. Identification of genes required for cytoplasmic localization in early *C. elegans* embryos. *Cell* **52**, 311–320 (1988).
- Watts, J. L., Morton, D. G., Bestman, J. & Kemphues, K. J. The *C. elegans par-4* gene encodes a putative serine–threonine kinase required for establishing embryonic asymmetry. *Development* **127**, 1467–1475 (2000).
- Guo, S. & Kemphues, K. J. *par-1*, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* **81**, 611–620 (1995).
- Shulman, J. M., Benton, R. & St Johnston, D. The *Drosophila* homolog of *C. elegans* PAR-1 organizes the oocyte cytoskeleton and directs oskar mRNA localization to the posterior pole. *Cell* **101**, 377–388 (2000).
- Tomanek, P. *et al.* A *Drosophila melanogaster* homologue of *Caenorhabditis elegans par-1* acts at an early step in embryonic-axis formation. *Nature Cell Biol.* **2**, 458–460 (2000).
- Hemminki, A. *et al.* A serine/threonine kinase gene defective in Peutz–Jeghers syndrome. *Nature* **391**, 184–187 (1998).
- Jenne, D. E. *et al.* Peutz–Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nature Genet.* **18**, 38–43 (1998).
- Collins, S. P., Reoma, J. L., Gamm, D. M. & Uhler, M. D. LKB1, a novel serine/threonine protein kinase and potential tumour suppressor, is phosphorylated by cAMP-dependent protein kinase (PKA) and prenylated *in vivo*. *Biochem. J.* **345**, 673–680 (2000).
- Sapkota, G. P. *et al.* Phosphorylation of the protein kinase mutated in Peutz–Jeghers cancer syndrome, LKB1/STK11, at Ser431 by p90^{RSK} and cAMP-dependent protein kinase, but not its farnesylation at Cys433, is essential for LKB1 to suppress cell growth. *J. Biol. Chem.* **276**, 19469–19482 (2001).
- van Eeden, F. & St Johnston, D. The polarisation of the anterior–posterior and dorsal–ventral axes during *Drosophila* oogenesis. *Curr. Opin. Genet. Dev.* **9**, 396–404 (1999).
- St Johnston, D., Beuchle, D. & Nüsslein-Volhard, C. *Staufén*, a gene required to localize maternal RNAs in the *Drosophila* egg. *Cell* **66**, 51–63 (1991).
- Clark, I., Giniger, E., Ruohola-Baker, H., Jan, L. Y. & Jan, Y. L. Transient posterior localization of a kinesin fusion protein reflects anteroposterior polarity of the *Drosophila* oocyte. *Curr. Biol.* **4**, 289–300 (1994).

13. Martin, S. G., Dobi, K. C. & St Johnston, D. A rapid method to map mutations in *Drosophila*. *Genome Biol.* **2**, research0036.1–12 (2001).
14. Benton, R., Palacios, I. M. & St Johnston, D. *Drosophila* 14–3–3/PAR-5 is an essential mediator of PAR-1 function in axis formation. *Dev. Cell* **3**, 659–671 (2002).
15. Spradling, A. *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez-Arias, A.) 1–70 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1993).
16. Huynh, J. R., Shulman, J. M., Benton, R. & St Johnston, D. PAR-1 is required for the maintenance of oocyte fate in *Drosophila*. *Development* **128**, 1201–1209 (2001).
17. Cox, D. N., Lu, B., Sun, T. Q., Williams, L. T. & Jan, Y. N. *Drosophila par-1* is required for oocyte differentiation and microtubule organization. *Curr. Biol.* **11**, 75–87 (2001).
18. de Cuevas, M. & Spradling, A. C. Morphogenesis of the *Drosophila* fusome and its implications for oocyte specification. *Development* **125**, 2781–2789 (1998).
19. Rorth, P. Gal4 in the *Drosophila* female germline. *Mech. Dev.* **78**, 113–118 (1998).
20. Lane, M. E. & Kalderon, D. RNA localization along the anteroposterior axis of the *Drosophila* oocyte requires PKA-mediated signal transduction to direct normal microtubule organization. *Genes Dev.* **8**, 2986–2995 (1994).
21. Hemminki, A. The molecular basis and clinical aspects of Peutz–Jeghers syndrome. *Cell. Mol. Life Sci.* **55**, 735–750 (1999).
22. Sanchez-Cespedes, M. *et al.* Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer Res.* **62**, 3659–3662 (2002).
23. Yoo, L. I., Chung, D. C. & Yuan, J. LKB1—a master tumour suppressor of the small intestine and beyond. *Nature Rev. Cancer* **2**, 529–535 (2002).
24. Bardeesy, N. *et al.* Loss of the Lkb1 tumour suppressor provokes intestinal polyposis but resistance to transformation. *Nature* **419**, 162–167 (2002).
25. Xu, T. & Rubin, G. M. Analysis of genetic mosaics in developing and adult *Drosophila* tissues. *Development* **117**, 1223–1237 (1993).
26. Chou, T. B. & Perrimon, N. The autosomal FLP–DFS technique for generating germline mosaics in *Drosophila melanogaster*. *Genetics* **144**, 1673–1679 (1996).
27. Sanson, B., White, P. & Vincent, J. P. Uncoupling cadherin-based adhesion from wingless signalling in *Drosophila*. *Nature* **383**, 627–630 (1996).
28. Wodarz, A., Hinz, U., Engelbert, M. & Knust, E. Expression of crumbs confers apical character on plasma membrane domains of ectodermal epithelia of *Drosophila*. *Cell* **82**, 67–76 (1995).
29. Sun, T. Q. *et al.* PAR-1 is a Dishevelled-associated kinase and a positive regulator of Wnt signalling. *Nature Cell Biol.* **3**, 628–636 (2001).
30. Theurkauf, W. *Drosophila melanogaster: Practical Uses in Cell and Molecular Biology* (eds Goldstein, L. & Fyrberg, E.) 489–506 (Academic, London, 1994).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Pal for fly injection; K. Litière for generating the kinesin– β -gal line; I. Palacios, B. Sanson, A. Wodarz and the Developmental Studies Hybridoma Bank for strains and antibodies; and J. Rouse, N. Lowe, J. Raff and R. Benton for advice and discussion. S.G.M. was supported by the Swiss National Science Foundation and D.StJ. by the Wellcome Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.StJ. (e-mail: ds139@mole.bio.cam.ac.uk).

Nicotinic acetylcholine receptor $\alpha 7$ subunit is an essential regulator of inflammation

Hong Wang*, Man Yu*, Mahendar Ochani*, Carol Ann Amella*, Mahira Tanovic*, Seenu Susarla*, Jian Hua Li*, Haichao Wang*, Huan Yang*, Luis Ulloa*, Yousef Al-Abed†, Christopher J. Czura* & Kevin J. Tracey*

* Laboratory of Biomedical Science, † Laboratory of Medicinal Chemistry, North Shore Long Island Jewish Research Institute, 350 Community Drive, Manhasset, New York 11030, USA

Excessive inflammation and tumour-necrosis factor (TNF) synthesis cause morbidity and mortality in diverse human diseases including endotoxaemia, sepsis, rheumatoid arthritis and inflammatory bowel disease^{1–4}. Highly conserved, endogenous mechanisms normally regulate the magnitude of innate immune responses and prevent excessive inflammation. The nervous system, through the vagus nerve, can inhibit significantly and

rapidly the release of macrophage TNF, and attenuate systemic inflammatory responses^{5–7}. This physiological mechanism, termed the 'cholinergic anti-inflammatory pathway'⁵ has major implications in immunology and in therapeutics; however, the identity of the essential macrophage acetylcholine-mediated (cholinergic) receptor that responds to vagus nerve signals was previously unknown. Here we report that the nicotinic acetylcholine receptor $\alpha 7$ subunit is required for acetylcholine inhibition of macrophage TNF release. Electrical stimulation of the vagus nerve inhibits TNF synthesis in wild-type mice, but fails to inhibit TNF synthesis in $\alpha 7$ -deficient mice. Thus, the nicotinic acetylcholine receptor $\alpha 7$ subunit is essential for inhibiting cytokine synthesis by the cholinergic anti-inflammatory pathway.

Nicotinic acetylcholine receptors are a family of ligand-gated, pentameric ion channels. In human, 16 different subunits ($\alpha 1$ – $\alpha 7$, $\alpha 9$ – $\alpha 10$, $\beta 1$ – $\beta 4$, δ , ϵ , γ) have been identified that form a large number of homo- and heteropentameric receptors with distinct structural and pharmacological properties^{8–10}. The main function of this receptor family is to transmit signals for the neurotransmitter acetylcholine at neuromuscular junctions and in the central and peripheral nervous systems^{8–12}. Our previous studies have indicated that acetylcholine inhibits the release of TNF and other cytokines through a post-transcriptional mechanism that is dependent on α -bungarotoxin-sensitive nicotinic receptors on primary human macrophages⁵, but the identity of the specific receptor subunit has remained unknown. As a first step towards identifying this macrophage receptor, primary human macrophages were labelled with fluorescein isothiocyanate (FITC)-tagged α -bungarotoxin, a peptide antagonist that binds to a subset of cholinergic receptors^{8,9}. Strong binding of α -bungarotoxin was observed on the macrophage surface (Fig. 1a). Nicotine pretreatment markedly reduced the

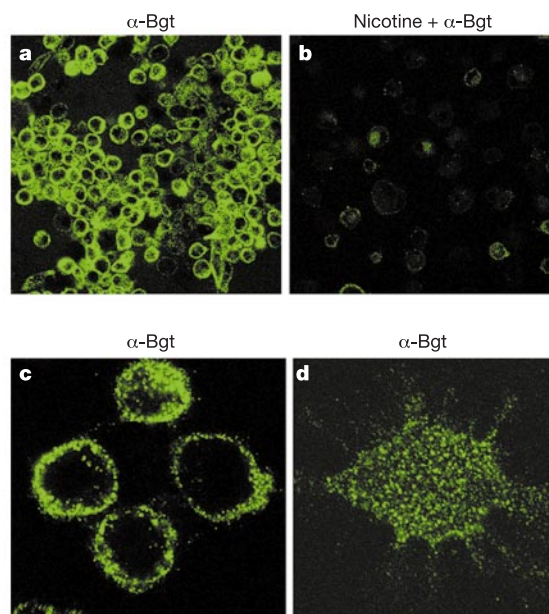


Figure 1 α -Bungarotoxin-binding nicotinic receptors are clustered on the surface of macrophages. Primary human macrophages were stained with fluorescein isothiocyanate (FITC)-labelled α -bungarotoxin (α -Bgt, $1.5 \mu\text{g ml}^{-1}$) and viewed by fluorescent confocal microscopy. **a**, Cells were stained with α -bungarotoxin alone. **b**, Nicotine was added to a final concentration of $500 \mu\text{mol}$ before addition of α -bungarotoxin. **c**, **d**, Higher magnification reveals receptor clusters. **c**, Focus planes are on the inside layers close to the middle (three lower cells) or close to the surface (upper cell) of cells. **d**, Focus plane is on the surface of the cell. Magnifications: **a**, **b**, $\times 50$; **c**, $\times 200$; **d**, $\times 450$.

intensity of binding (Fig. 1b). At neuromuscular junctions and at neuronal synapses, nicotinic receptors form receptor aggregates or clusters that facilitate fast signal transmission^{13–15}. Under high magnification, discrete clusters of α -bungarotoxin binding can be seen on the surface of macrophages. These clusters are particularly concentrated on the surface of the cell body (Fig. 1c, d).

$\alpha 1$, $\alpha 7$ and $\alpha 9$ had been described as potential α -bungarotoxin-binding nicotinic receptor subunits in mammalian cells^{8,9}. $\alpha 1$, together with $\beta 1$, δ and either ϵ (adult) or γ (fetal) subunits, can form heteropentameric nicotinic receptors that regulate muscle contraction; $\alpha 7$ and $\alpha 9$ can each form homopentameric nicotinic receptors^{8,9}. To determine whether these receptor subunits are expressed in macrophages, we isolated RNA from primary human macrophages differentiated *in vitro* from peripheral blood mononuclear cells (PBMCs), and performed polymerase chain reaction with reverse transcription (RT–PCR) analyses. To increase the sensitivity and specificity of the experiments, we conducted two rounds of PCR using nested primers specific to each subunit. The identities of the PCR products were confirmed by sequencing. Expression of $\alpha 1$, $\alpha 10$ (data not shown) and $\alpha 7$ (Fig. 2a) messenger RNA was detected in human macrophages derived from unrelated blood donors. The same RT–PCR strategy did not detect the expression of the δ -subunit, a necessary component of the $\alpha 1$ heteropentameric nicotinic acetylcholine receptor, or mRNA of the $\alpha 9$ subunit in human macrophages (data not shown).

The protein expression of $\alpha 1$ and $\alpha 7$ subunits was then examined by western blotting. An antibody specific for the nicotinic acetylcholine receptor $\alpha 7$ subunit (hereafter referred to as the $\alpha 7$ subunit) recognized a clear band with an apparent relative molecular mass of 55,000 (M_r 55K) (similar to the published molecular mass for $\alpha 7$ protein^{16,17}) from both differentiated primary macrophages and from undifferentiated PBMCs (data not shown). $\alpha 1$ protein expression was downregulated to undetectable levels during *in vitro* differentiation of PBMCs to macrophages (data not shown). To confirm that the positive signals in macrophages represented the $\alpha 7$ subunit that binds α -bungarotoxin, we used α -bungarotoxin-conjugated beads to pull-down proteins prepared

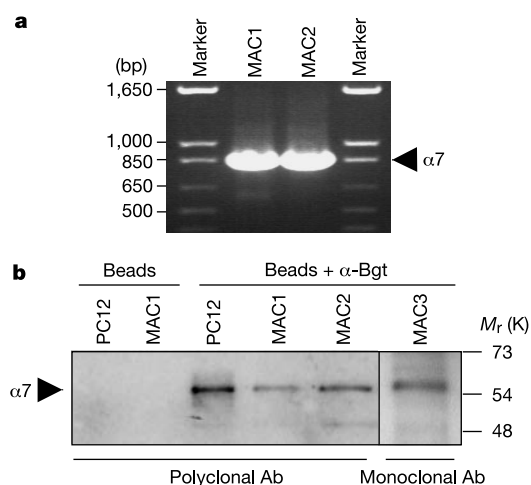


Figure 2 Messenger RNA and protein expression of nicotinic acetylcholine receptor $\alpha 7$ subunit in primary human macrophages. **a**, RT–PCR analysis. RT–PCR with primers specific for the $\alpha 7$ subunit generated an 843-base pair (bp) $\alpha 7$ band. PCR products were verified by sequencing (data not shown). MAC1 and MAC2 are macrophages derived from two unrelated donors. **b**, Western blots. Cell lysates from PC12 cells or human macrophages (MAC) were incubated with either control Sepharose beads or Sepharose beads conjugated with α -bungarotoxin. The bound proteins were analysed by $\alpha 7$ -specific polyclonal and monoclonal antibodies as indicated.

from either human macrophages or PC12 cells (rat pheochromocytoma cells, which express $\alpha 7$ homopentamer¹⁷). Retained proteins were analysed by western blotting using polyclonal or monoclonal $\alpha 7$ -specific antibodies that recognize both the human and rat $\alpha 7$ subunit (the human and rat $\alpha 7$ subunits contain the same number of amino acids and are 94% identical^{16,18}). We found that human macrophages express the α -bungarotoxin-binding $\alpha 7$ subunit with an apparent molecular mass similar to that of the $\alpha 7$ subunit isolated from PC12 cells (Fig. 2b). The identity of the macrophage $\alpha 7$ subunit was confirmed by cloning of the full-length complementary DNA of the macrophage-expressed $\alpha 7$ subunit by RT–PCR methods. The full-length cDNA of the $\alpha 7$ subunit in macrophages contains exons 1 to 10, identical to the $\alpha 7$ subunit expressed in neurons¹⁹. Together, these data identify the $\alpha 7$ subunit as the α -bungarotoxin-binding receptor expressed on the surface of human macrophages.

To study whether the $\alpha 7$ subunit is required for cholinergic inhibition of TNF release, we synthesized phosphorothioate antisense oligonucleotides surrounding the translation-initiation codon of the human $\alpha 7$ subunit gene. Antisense oligonucleotides to similar regions of the $\alpha 1$ and $\alpha 10$ subunit genes were synthesized as controls. Macrophages exposed to the antisense oligonucleotides specific for $\alpha 7$ (AS $\alpha 7$) were significantly less responsive to the TNF-inhibitory action of nicotine (Fig. 3a), and antisense oligonucleotides to the $\alpha 7$ subunit restored macrophage TNF release in the presence of nicotine. Exposure of macrophages to AS $\alpha 7$ did not stimulate TNF synthesis in the absence of lipopolysaccharide (LPS) and nicotine. Antisense oligonucleotides to $\alpha 1$ (AS $\alpha 1$) and $\alpha 10$

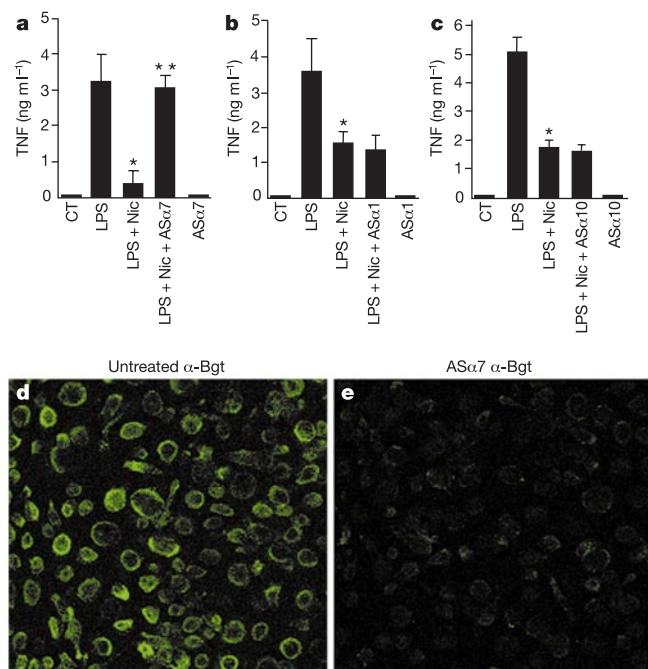


Figure 3 Antisense oligonucleotides to the $\alpha 7$ subunit inhibit the effect of nicotine on TNF release. **a–c**, TNF release from lipopolysaccharide (LPS)-stimulated primary human macrophages pretreated with antisense oligonucleotides to different subunits of nicotinic receptors. Where indicated, nicotine (Nic, 1 μ M) was added 5–10 min before LPS induction (100 ng ml⁻¹). TNF levels in the cell culture medium were determined by L929 assays. CT, control (unstimulated) macrophage cultures. AS $\alpha 7$, AS $\alpha 1$ and AS $\alpha 10$ are antisense oligonucleotides to $\alpha 7$, $\alpha 1$ and $\alpha 10$ subunits, respectively. Asterisk, $P < 0.05$ versus LPS; double asterisk, $P < 0.05$ versus LPS + Nic. **d, e**, FITC-labelled α -bungarotoxin staining of primary human macrophages treated (**e**) or untreated (**d**) with AS $\alpha 7$ and viewed by fluorescent confocal microscopy.

(AS α 10) subunits, under similar conditions, did not significantly change the effect of nicotine on LPS-induced TNF release (Fig. 3b, c), indicating that the suppression of TNF by nicotine is specific to the α 7 subunit. Further sets of antisense oligonucleotides to α 7, α 1 and α 10 subunits gave similar results (data not shown). Addition of AS α 7 to macrophage cultures decreased the surface binding of FITC-labelled α -bungarotoxin (Fig. 3d, e). Together these data indicate that the α 7 subunit is necessary for cholinergic inhibition of TNF release in macrophages.

Macrophages are an important source of TNF produced in response to bacterial endotoxin *in vivo*^{20,21}. To investigate whether the α 7 subunit is essential for the cholinergic anti-inflammatory pathway *in vivo*, we measured TNF production in mice deficient in the α 7 subunit gene originally generated by Orr-Urtreger and colleagues using genetic knockout technology²². In agreement with the original description of these mice²², mice lacking the α 7 subunit develop normally and show no gross anatomical defects^{22,23}. The serum TNF level in α 7-deficient mice after administration of endotoxin was significantly higher than wild-type endotoxaemic mice (wild-type serum TNF, 8.0 ± 2.4 ng ml⁻¹; α 7 knockout serum TNF, 18.1 ± 3.7 ng ml⁻¹, $P < 0.05$ (two-tailed *t*-test)) (Fig. 4a). TNF production in liver and spleen was also significantly higher in knockout mice (Fig. 4b, c), indicating a critical function of the α 7 subunit in the normal regulation of systemic inflammatory responses *in vivo*. Endotoxaemic α 7 knockout mice also produce significantly higher levels of interleukin (IL)-1 β (Fig. 4d) and IL-6 (Fig. 4e) as compared with wild-type mice. Macrophages derived from α 7 knockout mice were refractory to cholinergic agonists, and produced TNF normally in the presence of nicotine or acetylcholine (Table 1). Thus, α 7 subunit expression in macrophages is essential for cholinergic suppression of TNF.

To determine whether the α 7 subunit is required for vagus-nerve-dependent inhibition of systemic TNF, we applied electrical stimulation⁵ to the vagus nerve of endotoxaemic wild-type or

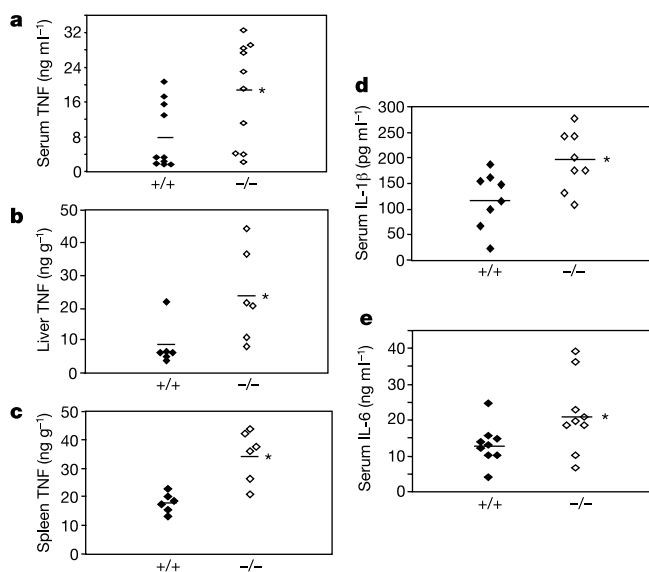


Figure 4 Increased cytokine production in α 7-subunit-deficient mice during endotoxaemia. α 7-deficient mice ($-/-$) or age- and sex-matched wild-type mice ($+/+$) were injected with LPS (0.1 mg kg⁻¹, intraperitoneally). Blood and organs were obtained either 1 h (for TNF) or 4 h (for IL-1 β and IL-6) after LPS administration. Levels of TNF, IL-1 β and IL-6 in serum or organs were measured with enzyme-linked immunosorbent assay (ELISA). **a**, TNF levels in serum; $n = 10$ per group; **b**, TNF levels in liver; $n = 6$ per group; **c**, TNF levels in spleen; $n = 6$ per group; **d**, IL-1 β levels in serum; $n = 8$ per group; **e**, IL-6 levels in serum; $n = 9$ per group. Asterisk, $P < 0.05$ versus wild-type controls. Horizontal line, the mean of the group.

Table 1 TNF production by wild-type and α 7-deficient peritoneal macrophages

	TNF (ng ml ⁻¹) wild type	TNF (ng ml ⁻¹) α 7 knockout
Control	0.004 ± 0.0005	0.004 ± 0.0005
LPS	16.8 ± 2.3	18.1 ± 4.9
LPS + nicotine (1μ M)	$5.2 \pm 0.9^*$	17.8 ± 0.6
LPS + nicotine (10μ M)	$7.3 \pm 1.0^*$	17.4 ± 2.9
LPS + Ach (1μ M)	10.3 ± 1.1	20.4 ± 3.8
LPS + Ach (10μ M)	$5.7 \pm 0.9^*$	21.4 ± 2.4

Thioglycollate-elicited peritoneal macrophages isolated from wild-type mice or nicotinic acetylcholine receptor α 7 subunit knockout mice were stimulated with LPS (100 ng ml⁻¹) for 4 h in culture. Control indicates unstimulated macrophage cultures. Where indicated, nicotine or acetylcholine (Ach) were added 5–10 min before LPS administration. TNF levels were measured by ELISA. Results are mean $\pm$ s.e.m.; $n = 8$ per group. Asterisk, $P < 0.05$ versus LPS.

α 7-deficient mice. Electrical stimulation of the vagus nerve significantly attenuated endotoxin-induced serum TNF levels in wild-type mice (Fig. 5). Vagus nerve stimulation using this protocol in α 7-deficient mice, however, failed to reduce serum TNF levels during endotoxaemia (Fig. 5). Thus, vagus nerve inhibition of TNF *in vivo* is dependent on the α 7 subunit.

These observations have several implications related both to understanding how the nervous system can regulate innate inflammatory responses in real time, and to designing experimental therapeutics that function via this physiological mechanism. Previous data indicate that the α 7 subunit forms homopentameric receptors that are involved in fast chemical signalling between cells^{8–10}. Neuronal α 7 receptors are highly permeable to calcium^{24,25}, and we have observed that nicotine induces transient calcium influx in macrophages (data not shown). Disruption of α 7 subunit expression *in vivo* significantly increases endotoxin-induced TNF release, indicating that the activity of this cholinergic anti-inflammatory pathway normally regulates the release of cytokines from macrophages and perhaps other cytokine-producing cells. It now seems that inactivation of this pathway can contribute to excessive systemic release of cytokines during endotoxaemia, and perhaps other states of infection or injury.

Deficiency of the α 7 subunit rendered the vagus nerve ineffective as a physiological pathway to inhibit TNF release, indicating that α 7 is essential for vagus nerve regulation of acute TNF release during the systemic inflammatory response to endotoxaemia. Thus, acetylcholine released from vagus nerve endings, or perhaps from other sources (for example, lymphocytes or epithelial cells) can specifically inhibit macrophage activation. It is quite possible that other primary immune cells (for example, mast cells, microglia cells, Kupffer cells, splenocytes) that produce cytokines during the acute, local response to infection or injury might express α 7 subunits, and

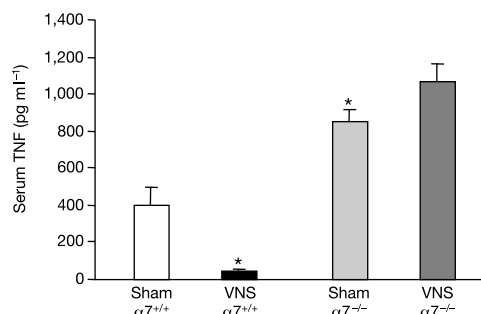


Figure 5 Vagus nerve stimulation does not inhibit TNF production in α 7-subunit-deficient mice. α 7-subunit-deficient mice ($-/-$) or age- and sex-matched wild-type mice ($+/+$) were subjected to either sham operation or vagus nerve stimulation (VNS, left vagus; 1 V, 2 ms, 1 Hz); blood was collected 2 h after LPS administration. Serum TNF levels were determined by ELISA. $n = 10$ (sham $\alpha 7^{+/+}$); $n = 11$ (VNS $\alpha 7^{+/+}$, sham $\alpha 7^{-/-}$, VNS $\alpha 7^{-/-}$). Asterisk, $P < 0.05$ versus sham $\alpha 7^{+/+}$.

that these cells might be sensitive to the anti-inflammatory effects of acetylcholine. Thus, there is significant potential for developing cholinergic agonists that target $\alpha 7$ subunits on peripheral immune cells for use as anti-inflammatory agents to inhibit release of TNF and other proinflammatory cytokines (for example, high mobility group box-1 protein)^{2,27}. Vagus nerve stimulators can enhance the anti-inflammatory activity of the cholinergic anti-inflammatory pathway in animals⁵⁻⁷; furthermore, vagus nerve stimulators have already been used safely in humans with seizure disorders. As TNF is already a clinically validated drug target for rheumatoid arthritis and Crohn's disease, it is now reasonable to consider the therapeutic potential for targeting the nicotinic acetylcholine receptor $\alpha 7$ subunit to inhibit TNF, either by direct pharmacological approaches, or through increasing activity in the vagus nerve. □

Methods

α -Bungarotoxin staining and confocal microscopy

Isolation and culture of human macrophages was performed as described previously⁵. Cells were differentiated for seven days in the presence of macrophage colony stimulating factor (MCSF; 2 ng ml⁻¹) in complete culture medium (RPMI 1640 with 10% heat-inactivated human serum). Differentiated macrophages were incubated with FITC-labelled α -bungarotoxin at 1.5 μ g ml⁻¹ (Sigma) in the cell culture medium at 4 °C for 15 min. Where indicated, nicotine was added to a final concentration of 500 μ M before the addition of α -bungarotoxin. Cells were washed three times with RPMI medium (Gibco) and then fixed for 15 min at room temperature in 4% paraformaldehyde-PBS solution (pH 7.2). After fixation, cells were washed with PBS once and mounted for viewing by fluorescent confocal microscopy.

RT-PCR

Total RNA was prepared from *in vitro* differentiated human macrophages using TRIzol reagent. Reverse transcription and the first round of PCR were performed using Titan One Tube RT-PCR Kit (Roche) according to the manufacturer's protocol. The second round of nested PCR was conducted using Promega PCR master mix. The PCR products from nested PCRs were run on agarose gels, recovered using the Gene Clean III Kit (BioLabs Inc.) and sent for sequencing to confirm the results. The primer sets for reverse transcription and the first round of PCR were: $\alpha 1$, sense primer 5'-CCAGACCTGAGCAACTTCATGG-3', antisense primer 5'-AATGAGTCGACCTGCAACACG-3'; $\alpha 7$, sense primer 5'-GACTGTTCCGTTCCAGATGG-3', antisense primer 5'-ACGAAGTTGGGAGCCGACA TCA-3'; $\alpha 9$, sense primer 5'-CGAGATCAGTACGATGGCCTAG-3', antisense primer 5'-TCTGTGACTAATCCGCTCTTGC-3'. The primer sets for nested PCRs were: $\alpha 1$, sense primer 5'-ATCACCTA-CCACTTCGTCATGC-3', antisense primer 5'-GTATGTGGTCCA TCACCATTGC-3'; $\alpha 7$, sense primer 5'-CCCGCAAGAGGAGTGAAAGGT-3', antisense primer 5'-TGCAGATGATGGTGAAGACC-3'; $\alpha 9$, sense primer 5'-AGAGCCT GTGAACACCAATGTGG-3', antisense primer 5'-ATGACTTTCGCCACCTTCTTCC-3'. For cloning of the full-length $\alpha 7$ cDNA, the following primers were used: 5'-AGGTGC CTCTGTGGCGCA-3' with 5'-GACTACTCAGTGGCCCTG-3'; 5'-CGACGCGGAGAC GTGGAG-3' with 5'-GGTACGGATGTGCCAAGGAGT-3'; 5'-CAAGGATCCGGACTC AACATGCGCTGCTCG-3' with 5'-CGGCTCGAGTCACCAAGTGTGGTTACGCAA GTC-3'.

Western blotting and α -bungarotoxin pull-down assay

Cell lysates were prepared by incubating PC12 cells or primary human macrophages with lysis buffer (150 mM NaCl, 5 mM EDTA, 50 mM Tris, pH 7.4, 0.02% NaN₃, 1% Triton X-100 and protease inhibitor cocktail) on ice for 90 min. Equal amounts of total protein were loaded on SDS-polyacrylamide gel electrophoresis (PAGE) for western blotting with either $\alpha 7$ -specific antibody (Santa Cruz sc-1447) or $\alpha 1$ monoclonal antibody (Oncogene). For the α -bungarotoxin pull-down assay, α -bungarotoxin (Sigma) was conjugated to CNBr-activated Sepharose beads (Pharmacia) and then incubated with cell lysates at 4 °C overnight. The beads and bound proteins were washed four times with lysis buffer and analysed by western blotting with $\alpha 7$ -specific antibodies (polyclonal, Santa Cruz H-302; monoclonal, Sigma M-220).

Antisense oligonucleotide experiments

Phosphorothioate antisense oligonucleotides were synthesized and purified by Genosys. The sequences for the oligonucleotides were: AS $\alpha 7$, 5'-GCAGCGCATGTTGAGTCCCG-3'; AS $\alpha 1$, 5'-GGGCTCCATGGGCTACCGGA-3'; AS $\alpha 10$, 5'-CCCCATGGCCCTGGCACTGC-3'. These sequences cover the divergent translation-initiation regions of $\alpha 7$, $\alpha 1$ and $\alpha 10$ genes. Delivery of the antisense oligonucleotides was carried out as in ref. 26 at 1 μ M concentration of the oligonucleotides for 24 h. For cell culture experiments, the oligonucleotide-pretreated macrophage cultures were washed with fresh medium and stimulated with 100 ng ml⁻¹ LPS with or without nicotine (1 μ M, added 5–10 min before LPS). At 4 h after LPS administration, the amounts of TNF released were measured by L299 assay and then verified by TNF enzyme-linked immunosorbent assay (ELISA). For α -bungarotoxin staining, pretreated cells were washed and processed for FITC- α -bungarotoxin staining as described above. Nicotine and other nicotinic acetylcholine receptor $\alpha 7$ subunit agonists also significantly inhibit LPS-induced TNF release in the murine macrophage-like cell line RAW264.7 (data not shown).

$\alpha 7$ -deficient mice

Mice deficient for the $\alpha 7$ nicotinic receptor (C57BL/6 background), and wild-type littermates were purchased from The Jackson Laboratory (B6.129S7-Chrna^{7ml1Bay}, number 003232). Breeding of homozygous knockout mice or wild-type mice was established to obtain progenies. Male or female mice about 8–12 weeks old (together with age- and sex-matched wild-type controls) were used for endotoxin experiments. Mice were weighed individually and 0.1 mg kg⁻¹ LPS was administered accordingly (intraperitoneal injection). For TNF experiments, blood, liver and spleen were collected 1 h after LPS administration. For IL-1 β and IL-6 experiments, blood samples were collected 4 h after LPS administration. The amounts of TNF, IL-1 β and IL-6 were measured by ELISA. We confirmed the genotypes of the mice by genomic PCR strategies. Peritoneal macrophages were isolated 48 h after intraperitoneal administration of thioglycollate (9%) from $\alpha 7$ knockout and wild-type male and female mice at 8 weeks of age ($n = 8$ per group). Macrophages were pooled for each group and cultured overnight. Nicotine and acetylcholine were added 5–10 min before LPS administration (100 ng ml⁻¹). Pyridostigmine bromide (100 μ M) was added with acetylcholine. Four hours after LPS administration, we measured TNF levels by ELISA.

Vagus nerve stimulation

$\alpha 7$ -deficient mice (C57BL/6 background, male and female) and age- and sex-matched wild-type C57BL/6 mice were anaesthetized with ketamine (100 mg kg⁻¹, intramuscularly) and xylazine (10 mg kg⁻¹, intramuscularly). Mice were either subjected to sham operation or vagus nerve stimulation (left cervical vagus, 1 V, 2 ms, 1 Hz) with an electrical stimulation module (STM100A, Harvard Apparatus). Stimulation was performed for 20 min (10 min before and 10 min after LPS administration). LPS was given at a lethal dose (75 mg kg⁻¹, intraperitoneally). Blood was collected 2 h after LPS administration. TNF levels were measured by ELISA.

Statistical analysis

Statistical analysis was performed using a two-tailed *t*-test where indicated; $P < 0.05$ is considered significant. Experiments were performed in duplicate or triplicate; for *in vivo* and *ex vivo* experiments, *n* refers to the number of animals under each condition.

Received 31 October; accepted 2 December 2002; doi:10.1038/nature01339.

Published online 22 December 2002.

- Tracey, K. J. *et al.* Shock and tissue injury induced by recombinant human cachectin. *Science* **234**, 470–474 (1986).
- Wang, H. *et al.* HMGI-1 as a late mediator of endotoxin lethality in mice. *Science* **285**, 248–251 (1999).
- Tracey, K. J. *et al.* Anti-cachectin/TNF monoclonal antibodies prevent septic shock during lethal bacteraemia. *Nature* **330**, 662–664 (1987).
- Tracey, K. J. & Cerami, A. Tumor necrosis factor: a pleiotropic cytokine and therapeutic target. *Annu. Rev. Med.* **45**, 491–503 (1994).
- Borovikova, L. V. *et al.* Vagus nerve stimulation attenuates the systemic inflammatory response to endotoxin. *Nature* **405**, 458–462 (2000).
- Bernik, T. R. *et al.* Pharmacological stimulation of the cholinergic anti-inflammatory pathway. *J. Exp. Med.* **195**, 781–788 (2002).
- Tracey, K. J. The inflammatory reflex. *Nature* **420**, 853–859 (2002).
- Lindstrom, J. M. in *Hand Book of Receptors and Channels: Ligand- and Voltage-Gated Ion Channels* (ed. North, A.) 153–175 (CRC Press, Boca Raton, Florida, 1995).
- Leonard, S. & Bertrand, D. Neuronal nicotinic receptors: from structure to function. *Nicotine Tob. Res.* **3**, 203–223 (2001).
- Le Novère, N. & Changeux, J.-P. Molecular evolution of the nicotinic acetylcholine receptor: an example of multigene family in excitable cells. *J. Mol. Evol.* **40**, 155–172 (1995).
- Marubio, L. M. & Changeux, J.-P. Nicotinic acetylcholine receptor knockout mice as animal models for studying receptor function. *Eur. J. Pharmacol.* **393**, 113–121 (2000).
- Steinlein, O. New functions for nicotine acetylcholine receptors? *Behav. Brain Res.* **95**, 31–35 (1998).
- Lin, W. *et al.* Distinct roles of nerve and muscle in postsynaptic differentiation of the neuromuscular synapse. *Nature* **410**, 1057–1064 (2001).
- Feng, G., Steinbach, J. H. & Sanes, J. R. Rapsyn clusters neuronal acetylcholine receptors but is inessential for formation of an interneuronal cholinergic synapse. *J. Neurosci.* **18**, 4166–4176 (1998).
- Shoop, R. D., Yamada, N. & Berg, D. K. Cytoskeletal links of neuronal acetylcholine receptors containing $\alpha 7$ subunits. *J. Neurosci.* **20**, 4021–4029 (2000).
- Peng, X. *et al.* Human $\alpha 7$ acetylcholine receptor: cloning of the $\alpha 7$ subunit from the SH-SY5Y cell line and determination of pharmacological properties of native receptors and functional $\alpha 7$ homomers expressed in *Xenopus* oocytes. *Mol. Pharmacol.* **45**, 546–554 (1994).
- Drisdell, R. C. & Green, W. N. Neuronal α -bungarotoxin receptors are $\alpha 7$ subunit homomers. *J. Neurosci.* **20**, 133–139 (2000).
- Seguela, P. *et al.* Molecular cloning, functional properties, and distribution of rat brain $\alpha 7$: a nicotinic cation channel highly permeable to calcium. *J. Neurosci.* **13**, 596–604 (1993).
- Gault, J. *et al.* Genomic organization and partial duplication of the human $\alpha 7$ neuronal nicotinic acetylcholine receptor gene (CHRNA7). *Genomics* **52**, 173–185 (1998).
- Bianchi, M. *et al.* An inhibitor of macrophage arginine transport and nitric oxide production (CNI-1493) prevents acute inflammation and endotoxin lethality. *Mol. Med.* **1**, 254–266 (1995).
- Kamins, N. H., Hunt, J., Gamelli, R. L. & Filkins, J. P. Partial hepatectomy reduces the endotoxin-induced peak circulating level of tumour necrosis factor in rats. *Shock* **5**, 385–388 (1996).
- Orr-Urtreger, A. *et al.* Mice deficient in the $\alpha 7$ neuronal nicotinic acetylcholine receptor lack α -bungarotoxin binding sites and hippocampal fast nicotinic currents. *J. Neurosci.* **17**, 9165–9171 (1997).
- Franceschini, D. *et al.* Altered baroreflex responses in $\alpha 7$ deficient mice. *Behav. Brain Res.* **113**, 3–10 (2000).
- Vijayaraghavan, S. *et al.* Nicotinic receptors that bind α -bungarotoxin on neurons raise intracellular free Ca²⁺. *Neuron* **8**, 353–362 (1992).

25. Shoop, R. D. *et al.* Synaptically driven calcium transients via nicotinic receptors on somatic spines. *J. Neurosci.* **21**, 771–781 (2001).
26. Cohen, P. S. *et al.* The critical role of p38 MAP kinase in T cell HIV-1 replication. *Mol. Med.* **3**, 339–346 (1997).
27. Andersson, U. *et al.* High mobility group 1 protein (HMG-1) stimulates proinflammatory cytokine synthesis in human monocytes. *J. Exp. Med.* **192**, 565–570 (2000).

Acknowledgements This work was supported in part by the National Institutes of Health, the National Institute of General Medical Sciences, and the Defense Advanced Research Projects Agency.

Competing interests statement The authors declare competing financial interests: details accompany the paper on *Nature's* website (► <http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to K.J.T. (e-mail: kjtracey@sprynet.com).

Activation of human CD4⁺ cells with CD3 and CD46 induces a T-regulatory cell 1 phenotype

Claudia Kemper^{*}, Andrew C. Chan[†], Jonathan M. Green[‡], Kelly A. Brett^{*§}, Kenneth M. Murphy^{||} & John P. Atkinson^{*¶}

^{*} Division of Rheumatology, Washington University School of Medicine, St Louis, Missouri 63110, USA

[†] Genentech, Inc., San Francisco, California 94080, USA

[‡] Department of Internal Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA

[§] Graduate Program in Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA

^{||} Division of Pathology and Immunology, Washington University School of Medicine, and Howard Hughes Medical Institute, St Louis, Missouri 63110, USA

The immune system must distinguish not only between self and non-self, but also between innocuous and pathological foreign antigens to prevent unnecessary or self-destructive immune responses. Unresponsiveness to harmless antigens is established through central and peripheral processes¹. Whereas clonal deletion and anergy are mechanisms of peripheral tolerance^{2,3}, active suppression by T-regulatory 1 (Tr1) cells has emerged as an essential factor in the control of autoreactive cells⁴. Tr1 cells are CD4⁺ T lymphocytes that are defined by their production of interleukin 10 (IL-10)⁵ and suppression of T-helper cells⁶; however, the physiological conditions underlying Tr1 differentiation are unknown. Here we show that co-engagement of CD3 and the complement regulator CD46 in the presence of IL-2 induces a Tr1-specific cytokine phenotype in human CD4⁺ T cells. These CD3/CD46-stimulated IL-10-producing CD4⁺ cells proliferate strongly, suppress activation of bystander T cells and acquire a memory phenotype. Our findings identify an endogenous receptor-mediated event that drives Tr1 differentiation and suggest that the complement system has a previously unappreciated role in T-cell-mediated immunity and tolerance.

Control of self-reactive T lymphocytes by regulatory T cells has been proposed to mediate anergy and thereby peripheral tolerance and prevention of autoimmunity⁷. Such cells suppress immune responses through either direct cell–cell interactions or the release of inhibitory cytokines such as IL-10 and transforming growth factor- β (TGF- β)^{8–11}. The differentiation of CD4⁺ lymphocytes into Tr1 cells is poorly defined, in part because of difficulties in inducing and culturing such cells. Only after the stimulation of

human CD4⁺ T cells with allogeneic monocytes or murine CD4⁺ T lymphocytes with antigen were some Tr1 clones generated; however, these cells had weak proliferative capacities and required the presence of IL-10 (ref. 5). Tr1 cell differentiation has also been induced by incubating CD4⁺ T cells with dexamethasone and vitamin D3 (ref. 12).

Defining the stimuli that drive differentiation of these cells would significantly enhance our understanding of their ontogeny and capabilities. Here we identify the complement regulatory protein CD46 as a physiological inducer of Tr1 cell development. CD46 is a widely expressed transmembrane glycoprotein that inhibits complement activation on host cells^{13,14}. CD46 is also used as a receptor by several human pathogens, including measles virus, herpesvirus 6 and two groups of pathogenic bacteria^{15–19}. Crosslinking CD46 through its physiological or pathogenic ligands initiates signalling events in several types of human cell^{20–23} and in cells derived from CD46 transgenic mice²⁴.

We stimulated purified human CD4⁺ lymphocytes with immobilized monoclonal antibodies and assessed their cytokine profile (Fig. 1). Cells stimulated with CD3 and CD46 (CD3/CD46) synthesized small amounts of IL-2, large amounts of IL-10 and intermediate amounts of TGF- β . By contrast, CD3/CD28 activation induced large amounts of IL-2, no IL-10 and barely detectable amounts of TGF- β . Neither stimulatory condition induced IL-4, but both led to the expression of similar quantities of IL-12 and interferon- γ (IFN- γ). Thus, CD3/CD46-stimulated cells have a cytokine profile that is distinct from that induced by stimulation with antibodies to CD3 and CD28 but similar to that of Tr1 cells⁵. IL-12 in the cultures suggests the presence of antigen-presenting

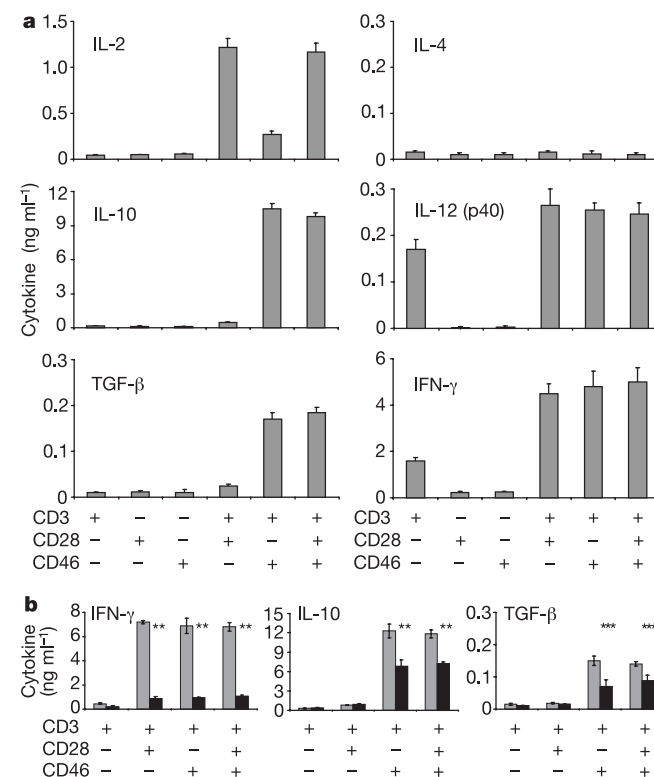


Figure 1 CD3/CD46 stimulation induces IL-10 production in human peripheral blood CD4⁺ T lymphocytes. **a**, Cytokine profile of stimulated T cells. Supernatants were collected at day 3. The same cytokine profile was observed with a second monoclonal antibody to CD46 (TRA-2-10). **b**, Effect of IL-12 neutralization. Cells were incubated with the indicated monoclonal antibodies in the presence (black bars) or absence (grey bars) of a neutralizing monoclonal antibody to IL-12. In all figures, values shown are the mean $\pm$ s.d. of at least three experiments. ** P < 0.005, *** P < 0.01.

¶ Present address: School of Medicine, 660 South Euclid Avenue, Campus Box 8045, St Louis, Missouri 63110, USA.

cells (APCs). Consistent with this observation, neutralization of this IL-12 completely blocked the synthesis of IFN- γ and reduced IL-10 and TGF- β production by about 40% (Fig. 1b). Thus, CD3/CD46 activation seems to be sufficient for induction of IL-10, but its production can be modulated by factors from other cells.

Purified CD4⁺ T lymphocytes were sorted for naive (CD45RA⁺CD45RO⁻) and memory (CD45RA⁻CD45RO⁺) cells. Naive and memory T cells constitute 20–40% of the CD4⁺ population of most donors. We also analysed the cell population with a CD45RA⁺CD45RO⁺ phenotype, because these cells also constitute about a third of peripheral blood CD4⁺ T cells. These three T-cell subpopulations expressed nearly the same quantities of CD3 and CD46 on their surface, as determined by fluorescence-activated cell sorting (FACS; data not shown). The sorted cell populations were activated in the presence or absence of IL-2 (Fig. 2). Memory CD4⁺ T cells produced IL-2 on CD3/CD28 activation but did not express IL-10 under any of the activation conditions (data not shown).

By contrast, CD45RA⁺CD45RO⁺ and naive CD4⁺ T cells both synthesized IL-10 after activation with CD3/CD28/CD46 or CD3/CD46 plus IL-2 (results with naive CD4⁺ T cells after primary and secondary activation are shown in Supplementary Fig. 1). The CD45RA⁺CD45RO⁺ subpopulation (high-responding cells) responded more robustly (Fig. 2a, c) in that they produced IL-10 earlier and in greater quantities (about 3–8-fold more than naive T cells). Neither naive (Supplementary Fig. 1a) nor high-responding (Fig. 2a) cells produced IL-4 or IL-12; however, both cell popu-

lations secreted moderate quantities of IFN- γ . Expression of TGF- β was more variable, with small quantities detectable in four out of seven donors. The ability of CD46 to induce IL-10 in isolated CD8⁺ T cells, natural killer cells, dendritic cells, monocytes or B lymphocytes was also examined. None of these cell populations produced IL-10 after activation with CD3/CD28/CD46 or CD3/CD46 plus IL-2 (data not shown).

The above data established that primary activation of naive or high-responding CD4⁺ T cells with CD3/CD46 in the presence of IL-2 generates a Tr1 phenotype. To determine whether these cells are then committed to maintain this phenotype, we analysed the properties of these CD4⁺ T cells on secondary stimulation. Sorted, high-responding CD4⁺ T cells were initially stimulated for 3 d (Fig. 2b) and subsequently expanded for 6 d in media supplemented with IL-2. The cells were then subjected to secondary stimulation, and IL-10 production was measured after 2 d. High-responding CD4⁺ T cells, first activated with either CD3/CD28/CD46 or CD3/CD46 plus IL-2, produced large amounts of IL-10 on secondary stimulation (Fig. 2b). Supplementation with IL-2 during secondary stimulation enhanced IL-10 secretion by about 30% (Fig. 2b). The addition of a neutralizing monoclonal antibody to IL-2 abrogated IL-10 production (data not shown), showing that IL-10 production is dependent on IL-2 for both primary and secondary CD46-activated responses.

By contrast, CD4⁺ T cells initially activated without CD46 crosslinking produced minimal amounts of IL-10 on secondary

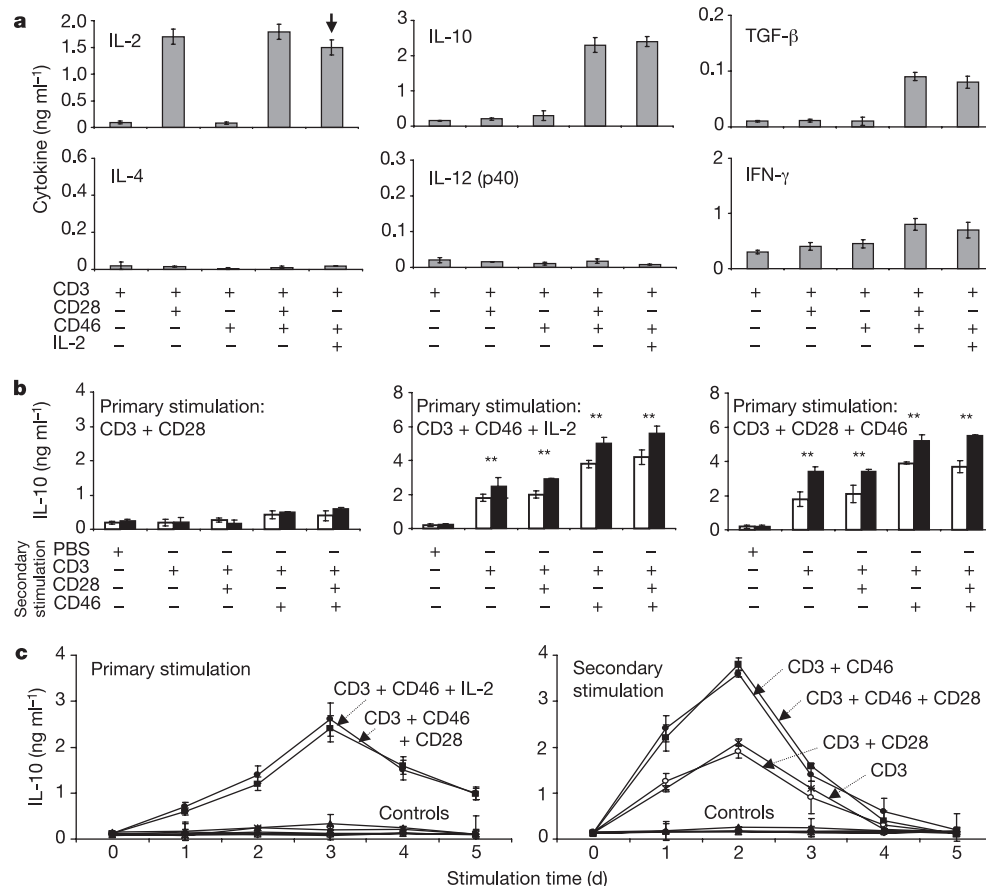


Figure 2 Sorted CD3⁺CD4⁺CD45RA⁺CD45RO⁺ T cells respond to primary and secondary activation with IL-10 production. **a**, Cytokine profile of primary-stimulated cells. Cells were stimulated for 3 d (with or without IL-2). Arrow indicates the detection of added recombinant IL-2. **b**, Production of IL-10 on secondary stimulation. After primary stimulation for 3 d, cells were expanded and subjected to secondary stimulation in the presence (black bars) or absence of IL-2 (white bars). IL-10 secretion was measured at

day 2. **c**, Kinetics of IL-10 production. Left, IL-10 synthesis after primary stimulation. Controls, incubation with monoclonal antibodies to CD3 plus CD28, CD3 plus CD46, CD3 plus isotype control monoclonal antibody, or CD3 plus CD46 plus isotype control. Right, after primary activation with CD3/CD46 plus IL-2, cells were expanded and then subjected to secondary stimulation with the indicated monoclonal antibodies. Controls, incubations with monoclonal antibodies to CD28 or CD46 or with IL-2 alone.

stimulation with CD3/CD46. Therefore, primary stimulation through CD46 is required for IL-10 production in restimulated cells. Notably, with a secondary stimulation, the activation of CD3 alone became sufficient for the induction of IL-10 synthesis (Fig. 2b,c), indicating transition to a memory IL-10-producing phenotype. In addition, IL-10 secretion during secondary activation began and peaked earlier, and was of greater magnitude than that induced by primary activation. Restimulated cells did not produce IL-4 or IL-12, but synthesized small amounts of IL-2 and moderate amounts of IFN- γ (data not shown). In this same set of experiments, naive CD4⁺ T cells responded in a comparable fashion (Supplementary Fig. 1).

Highly purified high-responding or naive CD4⁺ T cells, after primary activation with CD3/CD46, required the presence of a monoclonal antibody to CD28 or the addition of exogenous IL-2 to produce IL-10. Neutralization of IL-2 abrogated IL-10 production (Fig. 3a). Because CD3/CD46 stimulation is not sufficient to induce production of IL-2, the IL-10 secretion observed in non-sorted CD4⁺ T cells on stimulation with CD3/CD46 alone (Fig. 1a) is

probably due to engagement of CD28 by endogenous B7 expression of APCs in the culture. The primary role of CD28 in the induction of IL-10 may be to provide IL-2.

Because Tr1 cells are characterized by release of IL-10 without concurrent production of IL-2, we examined whether naive and high-responding CD4⁺ T cells synthesize IL-10 and/or IL-2 by intracellular cytokine staining (Fig. 3b). Cell cultures activated with CD3/CD28/CD46 for 12 h contained mostly non-overlapping populations of IL-10- and IL-2-producing cells. Activation of cells with CD3/CD46 alone did not lead to the production of either cytokine, whereas supplementation with recombinant IL-2 generated only a population of IL-10-expressing cells (Fig. 3b).

CD45RA, CD45RO and CD25 are cell-surface markers that assess the activation state of T cells, specifically their switch from naive (CD45RA⁺CD45RO⁻CD25⁻) to memory (CD45RA⁻CD45RO⁺CD25⁺) phenotype after CD3/CD28 activation. To determine whether CD3/CD46 activation induces similar changes, we analysed expression of these markers by sorted CD4⁺ T-cell populations (Fig. 3c, d). Indeed, CD46-activated cells did acquire a memory

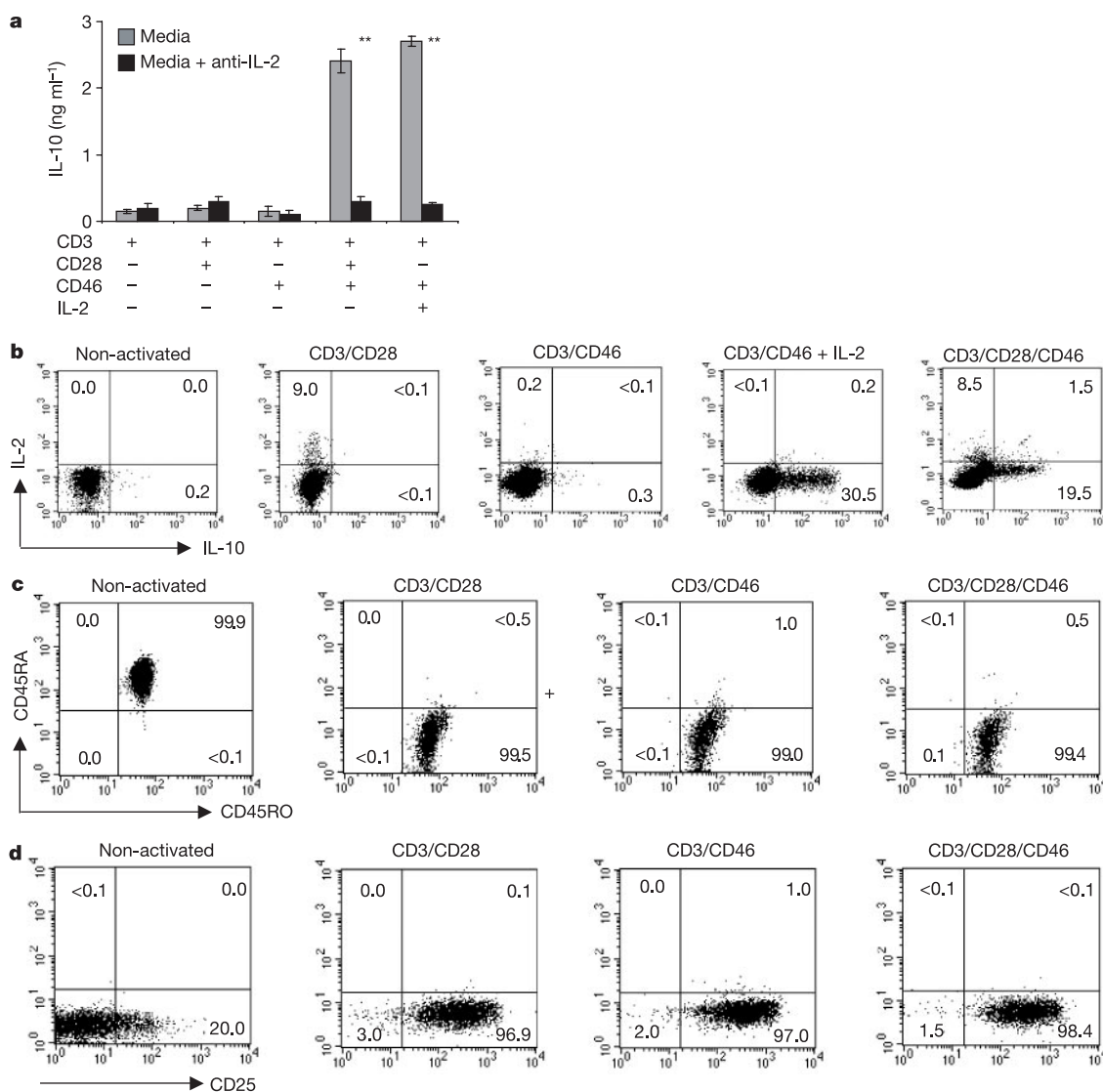


Figure 3 Characteristics of CD3/CD46-activated, sorted CD3⁺CD4⁺CD45RA⁺CD45RO⁺ T cells. **a**, CD3/CD46-induced IL-10 production is IL-2 dependent. T cells were incubated with the indicated immobilized monoclonal antibodies with or without a monoclonal antibody to IL-2. Supernatants were collected at day 3. **b**, IL-10-producing cells do not express IL-2. T cells were stimulated for 12 h and intracellular cytokine staining was

carried out. **c**, CD3/CD46-activated T cells acquire an CD45RA⁻CD45RO⁺ phenotype after primary stimulation. T cells were stimulated with the indicated immobilized monoclonal antibodies. Surface expression of CD45RA and CD45RO was monitored at day 3. **d**, CD3/CD46-activated T cells upregulate CD25 after primary stimulation. T cells were stimulated as indicated for 24 h and expression of CD25 was analysed.

phenotype, because the high-responding, as well as the naive, CD4⁺ T cells became nearly 100% CD45RA⁻CD45RO⁺CD25⁺. This phenotype was not altered after secondary stimulation (data not shown).

We next addressed whether the supernatant from CD3/CD46-activated T cells suppresses proliferation of bystander T cells. At day 3, supernatants from the activated cell populations were transferred to freshly purified CD4⁺ T cells. These cells were then activated with CD3/CD28 and proliferation was measured at day 7. The supernatants from CD4⁺ T cells activated by both CD3/CD46

plus IL-2 and CD3/CD28/CD46 inhibited proliferation of bystander CD4⁺ T cells (Fig. 4a, top). This inhibition was not observed when a neutralizing monoclonal antibody to IL-10 was added (Fig. 4a, bottom). Supernatants derived from sorted, naive CD4⁺ T cells, from non-sorted CD4⁺ T cells and from secondary-stimulated CD4⁺ T cells also inhibited proliferation of bystander T cells (data not shown). As expected, supernatants from CD3/CD28-activated CD4⁺ T cells did not inhibit proliferation of T cells.

Tr1 cells, derived *in vitro* by stimulation of human CD4⁺ cells in the presence of IL-10, proliferate poorly on antigenic stimulation^{5,11}. By contrast, activation of CD46 induces IL-10 (as shown above) and, as previously reported²², is a potent costimulator of CD4⁺ T-cell proliferation. CD3/CD46-activated CD4⁺ T cells showed a stronger and more sustained proliferative response as compared with CD3/CD28-activated cells (Fig. 4b). Specifically, CD28/CD3-activated cells ceased proliferating by day 7, whereas CD3/CD46-stimulated cells expanded until at least day 11.

The CD3/CD46-mediated induction of cell proliferation and of IL-10 production might be independently triggered events. To analyse whether IL-10-secreting cells are also proliferating, we stained activated cells intracellularly for IL-10 and for proliferating cell nuclear antigen (PCNA). Roughly 60% of the IL-10-secreting cells proliferated strongly, and this response was comparable to that of non-IL-10-producing, CD3/CD46-activated CD4⁺ T cells (Fig. 4c). Thus, CD3/CD46-activated cells proliferate strongly despite the presence of high amounts of IL-10, whereas CD3/CD28-activated cells cease proliferation in the presence of IL-10. A possible explanation for this observation might be that CD46 signalling downregulates IL-10 receptor expression. We therefore analysed IL-10 receptor expression of CD3/CD46-activated (IL-10-producing) and CD3/CD28-activated (IL-2-producing) cells. There was no difference in receptor expression (data not shown), suggesting that Tr1 cells evade the inhibitory effects of IL-10 by another mechanism.

To assess whether a physiological ligand of CD46 can also induce IL-10, we activated high-responding CD4⁺ T cells with CD3, CD28 and C3b dimers (in place of monoclonal antibodies to CD46). Activation with CD3/CD28/C3b dimers resulted in IL-10 secretion comparable to that of CD3/CD28/CD46 (Supplementary Fig. II), thus providing a mechanism for CD46 crosslinking by antigens coated with complement fragments during antigen presentation (Supplementary Fig. III).

In summary, signalling events mediated by CD46 activation induce the development of CD4⁺ T cells to a Tr1 phenotype. This suggests a role for CD46 in human T-cell regulation and establishes a link between the complement system and adaptive immunity^{25–27}. Tr1 cells are essential for maintaining peripheral tolerance and preventing autoimmunity^{5,11}. They may also modulate the host's immune response to many pathogens²⁸. This putative role of CD46 in regulating cellular immune responses might also explain why several human pathogens have chosen CD46 as their receptor²⁴. Our identification of CD46 as a physiological stimulus for Tr1 cell differentiation and proliferation should facilitate further characterization of suppressor cell populations and might provide a therapeutic strategy for generating such cells to ameliorate T-cell-mediated pathology. □

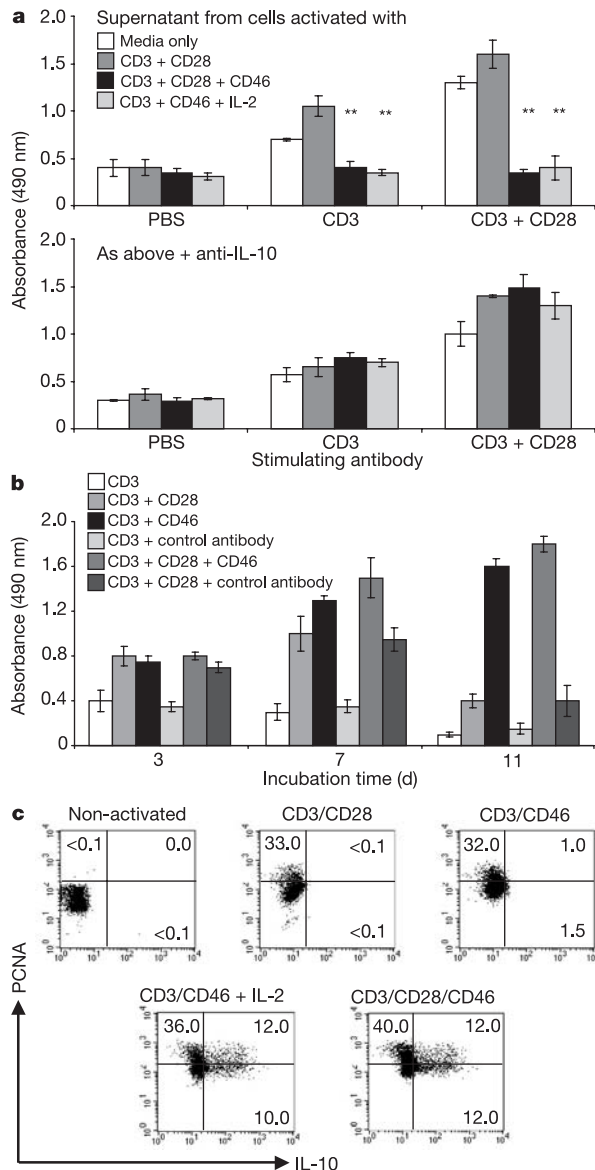


Figure 4 Suppressive and proliferative properties of CD3/CD46-activated CD4⁺ T cells. **a**, CD3/CD46-stimulated T cells inhibit the proliferation of bystander CD4⁺ T cells through IL-10. Top, sorted CD3⁺CD4⁺CD45RA⁺CD45RO⁺ T cells were stimulated for 3 d with the indicated antibodies. Supernatants were transferred to freshly purified CD4⁺ lymphocytes and the mixture was incubated with the indicated immobilized antibodies. Proliferation was measured at day 7. Bottom, as above but with the addition of a neutralizing monoclonal antibody to IL-10. **b**, CD46 induces a sustained proliferative response. Purified CD4⁺ T cells were activated with immobilized antibodies and proliferation was measured. **c**, CD3/CD46-stimulated IL-10-producing cells proliferate. Sorted CD3⁺CD4⁺CD45RA⁺CD45RO⁺ T cells were activated for 12 h with the indicated antibodies and stained for intracellular IL-10 and PCNA.

Methods

Media and antibodies

We obtained media and supplements from the tissue-culture facility at Washington University School of Medicine. Cells were maintained in RPMI medium with 10% fetal calf serum and 200 mM L-glutamine in the presence or absence of recombinant human IL-2 (BioSource International). The monoclonal antibody GB24 to human CD46 binds to complement control protein repeats 3 and 4 (ref. 29). The other monoclonal antibody to CD46 that we used, TRA-2-10, recognizes an epitope in the first repeat²⁹. The monoclonal antibodies to human CD3 (HIT3a) and CD28 (CD28.2), and the function-neutralizing monoclonal antibodies to human IL-10 (JES3-9D7), IL-2 (MQ1-17H12) and IL-12 (C8.6) were purchased from BD PharMingen. Fluorescein isothiocyanate (FITC)- or

phycoerythrin (PE)-labelled monoclonal antibodies to IL-10, IL-2 and IL-4 (BD PharMingen) and PCNA (Santa Cruz) were used for intracellular staining. We obtained the mouse nonspecific isotype-matched control monoclonal antibody (MOPC 31c, IgG1) from Sigma-Aldrich. The monoclonal antibodies to human CD14 (MoP9), CD16 (3G8), CD8 (Leu-2b), CD19 (4G7) and HLA-DR (L243), used for purifying CD4⁺ cells, were from Becton Dickinson. Monoclonal antibodies to human CD3 (HIT3a), CD4 (RPA-T4), CD45RA (HI100), CD45RO (UCHL1) and CD25 (2A3), labelled with FITC, PE, allophycocyanin or PerCP, were obtained from Becton Dickinson and used for cell sorting. C3b dimers were generated as described³⁰.

Purification and sorting of CD4⁺ lymphocytes

CD4⁺ lymphocytes were purified from whole blood of healthy human donors by separating the mononuclear cells through Ficoll-Paque Plus (Amersham Pharmacia). This population was then incubated with a mixture of IgG monoclonal antibodies to CD8, CD19, CD14, CD16 and HLA-DR to deplete the CD8⁺ T cells, B lymphocytes, monocytes/macrophages and natural killer cells and dendritic cells, respectively. After incubation with goat antibodies against mouse IgG coupled to magnetic microbeads (Miltenyi Biotech), CD4⁺ lymphocytes were obtained by passing the cell mixture over MiniMACS magnetic separation columns (Miltenyi Biotech) and collecting the CD4⁺ cells in the flow through. For experiments with non-sorted cells, only samples with >97% CD4⁺ cells were used. To isolate T-cell subpopulations from the enriched purified CD4⁺ T cells, we incubated cells with combinations of fluorochrome-labelled monoclonal antibodies to CD3, CD4, CD45RA and CD45RO, and then sorted them sterily using a MoFlo high performance cytometer (Cytomation). Blood from seven donors was used in this study.

Primary and secondary cell stimulation

In vitro primary stimulation was carried out in 96-well culture plates coated with monoclonal antibodies to CD3 (10 µg ml⁻¹), CD28 (5 µg ml⁻¹), CD46 (5 µg ml⁻¹) or a matched IgG1 isotype control antibody (5 µg ml⁻¹). The wells were washed and purified CD4⁺ lymphocytes (1.5–2.0 × 10⁵ cells per well) were added in 150 µl of culture medium. The plates were centrifuged at 200g for 1 min and incubated at 37°C in 5% CO₂. After 3 d of primary stimulation, cells were washed and expanded for 6 d in media supplemented with 40 units ml⁻¹ human IL-2. At day 7, cells were counted and subjected to secondary stimulation under conditions similar to those of the primary activation. We carried out all experiments at least three times and analysed each activation condition in triplicate. Statistical significance was determined using the paired Student's *t*-test.

Proliferation assays

We measured cell proliferation rates using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay (Promega). This colorimetric assay determines the number of living cells using an MTS tetrazolium compound (Owen's reagent). Initial experiments using this system yielded equivalent results to those obtained by [³H]thymidine incorporation. To analyse which cells among an activated cell population expanded, cells (1.5–2.0 × 10⁵ per well) were stimulated with immobilized monoclonal antibodies for 12, 24 or 48 h. Brefeldin A (Sigma-Aldrich) was added for the last 4–8 h of culture. After appropriate surface marker staining, cells were fixed and permeabilized for 10 min with 4% paraformaldehyde and 0.1% saponin (Sigma-Aldrich) in Hank's balanced salt solution and then stained for PCNA. Experiments were done at least three times and each condition was analysed in triplicate.

Cytokine analyses

CD4⁺ cells (1.5–2.0 × 10⁵ per well) were incubated for up to 5 d in 96-well plates coated with monoclonal antibodies. We assessed the secretion of IL-2, IL-4, IFN-γ, IL-10, TGF-β and IL-12 in the supernatants by enzyme-linked immunosorbent assay. IL-4 and IL-12 (p40 and p70) were measured using Biotrak Cellular Communication Assay (Amersham Pharmacia). IL-2, IFN-γ, IL-10 and TGF-β were analysed using EASIA kits (BioSource International). To assess the number of cytokine producing cells within a stimulated cell population, cells (1.5–2.0 × 10⁵ cells per well) were activated with immobilized monoclonal antibodies for 12, 24, 48 or 72 h. Brefeldin A was added for the last 4–8 h of culture. After appropriate surface marker staining, cells were stained for intracellular cytokines.

Received 7 August; accepted 15 November 2002; doi:10.1038/nature01315.

1. Van Parijs, A. & Abbas, A. K. Homeostasis and self-tolerance in the immune system: turning lymphocytes off. *Science* **280**, 243–248 (1998).
2. Kappler, J. W., Roehm, N. & Marrack, P. T cell tolerance by clonal elimination in the thymus. *Cell* **49**, 273–280 (1987).
3. Kisielow, P., Bluthmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4⁺8⁺ thymocytes. *Nature* **333**, 742–746 (1988).

4. Sakaguchi, S. Regulatory T cells: key controllers of immunologic self-tolerance. *Cell* **101**, 455–458 (2000).
5. Groux, H. *et al.* A CD4⁺ T-cell subset inhibits antigen-specific T-cell responses and prevents colitis. *Nature* **389**, 737–742 (1997).
6. Roncarolo, M. G. & Levings, M. K. The role of different subsets of T regulatory cells in controlling autoimmunity. *Curr. Opin. Immunol.* **12**, 676–683 (2000).
7. LeDourarin, N. *et al.* Evidence for a thymus-dependent form of tolerance that is not based on elimination or anergy of reactive T cells. *Immunol. Rev.* **149**, 35–53 (1996).
8. Shevach, E. M. Regulatory T cells in autoimmunity. *Annu. Rev. Immunol.* **18**, 423–449 (2000).
9. Shevach, E. M. S. M. R., Piccirillo, C. A. & Thornton, A. M. Control of T-cell activation by CD4⁺CD25⁺ suppressor T cells. *Immunol. Rev.* **182**, 58–67 (2001).
10. Chen, Y., Kuchroo, V. K., Inobe, J., Hafler, D. A. & Weiner, H. L. Regulatory T cell clones induced by oral tolerance: suppression of autoimmune encephalomyelitis. *Science* **265**, 1237–1240 (1994).
11. Roncarolo, M. G., Bacchetta, R., Bordignon, C., Narula, S. & Levings, M. K. Type 1 regulatory cells. *Immunol. Rev.* **182**, 68–79 (2001).
12. Barrat, F. J. *et al.* *In vitro* generation of interleukin 10-producing regulatory CD4⁺ T cells is induced by immunosuppressive drugs and inhibited by T helper type 1 (Th1)- and Th2-inducing cytokines. *J. Exp. Med.* **195**, 603–616 (2002).
13. Cole, J. L., Housley, G. A. Jr, Dykman, T. R., MacDermott, R. P. & Atkinson, J. P. Identification of an additional class of C3-binding membrane proteins of human peripheral blood leukocytes and cell lines. *Proc. Natl Acad. Sci. USA* **82**, 859–863 (1985).
14. Liszewski, M. K., Post, T. W. & Atkinson, J. P. Membrane cofactor protein (MCP or CD46): newest member of the regulators of complement activation gene cluster. *Annu. Rev. Immunol.* **9**, 431–455 (1991).
15. Dorig, R. E., Marciel, A., Chopra, A. & Richardson, C. D. The human CD46 molecule is a receptor for measles virus (Edmonston strain). *Cell* **75**, 295–305 (1993).
16. Nanche, D. *et al.* Human membrane cofactor protein (CD46) acts as a cellular receptor for measles virus. *J. Virol.* **67**, 6025–6032 (1993).
17. Santoro, F. *et al.* CD46 is a cellular receptor for human herpes virus 6. *Cell* **99**, 817–827 (1999).
18. Okada, N., Liszewski, M. K., Atkinson, J. P. & Caparon, M. Membrane cofactor protein (MCP; CD46) is a keratinocyte receptor for the M protein of group A streptococcus. *Proc. Natl Acad. Sci. USA* **92**, 2489–2493 (1995).
19. Kallstrom, H., Liszewski, M. K., Atkinson, J. P. & Jonsson, A.-B. Membrane cofactor protein (MCP or CD46) is a cellular pilus receptor for pathogenic *Neisseria*. *Mol. Microbiol.* **25**, 639–647 (1997).
20. Karp, C. L. *et al.* Mechanism of suppression of cell-mediated immunity by measles virus. *Science* **273**, 228–231 (1996).
21. Kallstrom, H., Islam, M. S., Berggren, P. O. & Jonsson, A. B. Cell signaling by the type IV pili of pathogenic *Neisseria*. *J. Biol. Chem.* **273**, 21777–21782 (1998).
22. Astier, A., Trescol-Biemont, M.-C., Azocar, O., Lamouille, B. & Rabourdin-Combe, C. Cutting edge: CD46, a new costimulatory molecular for T cells, that induces p120CBL and LAT phosphorylation. *J. Immunol.* **164**, 6091–6095 (2000).
23. Wang, G., Liszewski, M. K., Chan, A. C. & Atkinson, J. P. Membrane cofactor protein (MCP; CD46): isoform-specific tyrosine phosphorylation. *J. Immunol.* **164**, 1839–1846 (2000).
24. Marie, J. *et al.* Linking innate and acquired immunity: divergent role of CD46 cytoplasmic domains in T cell-induced inflammation. *Nature Immunol.* **3**, 659–666 (2002).
25. Nielsen, C. H., Fischer, E. M. & Leslie, R. G. The role of complement in the acquired immune response. *Immunology* **100**, 4–12 (2000).
26. Dempsey, P. W., Allison, M. E., Akkaraju, S., Goodnow, C. C. & Fearon, D. T. C3d of complement as a molecular adjuvant: bridging innate and acquired immunity. *Science* **271**, 348–350 (1996).
27. Prodeus, A. P. *et al.* A critical role for complement in maintenance of self-tolerance. *Immunity* **9**, 721–731 (1998).
28. McGuirk, P. & Mills, K. H. G. Pathogen-specific regulatory T cells provoke a shift in the Th1/Th2 paradigm in immunity to infectious diseases. *Trends. Immunol.* **23**, 450–455 (2002).
29. Liszewski, M. K. *et al.* Dissecting sites important for complement regulatory activity in membrane cofactor protein (MCP; CD46). *J. Biol. Chem.* **275**, 37692–37701 (2001).
30. Krych-Goldberg, M. *et al.* Decay accelerating activity of complement receptor type 1 (CD35). *J. Biol. Chem.* **274**, 31160–31168 (1999).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank H. Molina, K. Liszewski, L. S. Wen and M. Denny for comments on the manuscript, and the immunology community at Washington University School of Medicine for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P.A. (e-mail: jatkins@im.wustl.edu).

Automation information

Sample tracking, sample preparation and other methods for automation.

Genesis FE500 Workcell

Tecan www.tecan.com

Added flexibility for this system.

Tecan has introduced several new features for the Genesis FE500 Workcell. This is a fully automated platform designed to meet the sample processing requirements of clinical laboratories. The new features follow on from requests from customers who use the Workcell for a front-end (pre-analytical) automation applications. The new system is modular, offering different options for the individual processing steps including specimen sorting; centrifugation; decapping; secondary tube labelling and aliquoting; serum volume verification; barcode labelling; and placing tubes in sample racks that originate from a wide variety of manufacturers. An improved tube inspection unit scans the whole length of the tube to measure the exact liquid volume. Since detection is possible through the tube label, no window is necessary on sample tubes and the complete tube may be covered by barcode labels (up to 3 layers). This feature is also available as an upgrade for existing users. New archiving options enable the full separation of sample from primary to secondary tubes.

Wizard MagneSil Tfx

Promega www.promega.com

High purity plasmid DNA for transfection

Aimed at industrial and academic laboratories that need a reliable and cost effective method for purifying transfection grade DNA, the MagneSil Tfx System has been designed for use in a multi-well format (8–96 samples). No centrifugation or vacuum steps are required making the system particularly suitable to full automation. A 96-well plate can be processed in just 45 minutes with a Beckman Biomek FX robotic workstation, so maximizing throughput. The protocol can also be easily adapted to suit



MagneSil Tfx: transfection DNA in 45 minutes.



Mix and match: the Genesis FE500 goes modular.

other robotic systems, if required. The Wizard MagneSil Tfx uses alkaline-SDS lysis to generate the bacterial lysate and incorporates MagneSilT paramagnetic particles for both lysate clearing and plasmid purification. The system also includes Promega's Endotoxin Removal Resin, which reduces the levels of contaminants such as chemical salts, RNA, protein and endotoxin, all of which can decrease DNA transfection efficiencies. The procedure is carried out in seven simple steps, and needs no manual intervention once an automatic protocol has been initiated.

IDM-LimsLinkCDS 3.1

Applied Biosystems
www.appliedbiosystems.com

Software upgrades

IDM-LimsLink provides an interface between Applied Biosystems' SQL LIMS software and all laboratory instruments. Version 3.1 introduces new features including an e-signature option to increase security and a runtime feature that enables users to collect data automatically. At the end of a run, a shutdown procedure ceases data collection and closes the worksheet. For chromatography data systems, IDM-LimsLinkCDS 3.1 provides a secure, bidirectional communication interface with SQL LIMS software. It is compatible with all standard chromatography data systems, including Waters Millennium 32, Agilent Chem-

Station, and PerkinElmer's TotalChrom and Turbochrom software.

Labworks ES

PerkinElmer www.perkinelmer.com

Data security with regulatory compliance

PerkinElmer's Labworks ES (Enhanced Security) software simplifies the process of securing critical laboratory data by providing a central point of authentication that can be integrated with different systems. New features include additional reporting tools, expanded data fields and barcode printing. The software meets 21 CFR Part 11 requirements and is suitable for use in environmental, chemical, petrochemical, water and wastewater, food and beverage applications.

Sample Tracker

Autoscribe www.autoscribe.co.uk

Alternative to LIMS

Sample Tracker can be used to register samples either singly or in batches, and will assign a unique number to each new sample or batch. Storage details such as room and freezer numbers are recorded, and relocation or removal of samples for testing can be tracked. Sample status is monitored either on screen or in a report, which contains information on sample movement history, expiry date review, storage room management and inventory. The system is compatible with Windows 98, NT, 2000 and XP.

i6000 staining system

BioGenex www.biogenex.com

Histopathology at a 'genomic' pace

BioGenex describe the i6000 automatic staining system as an 'all-in-one-all-at-once' instrument. Capacity is for 60 slides and 60 reagents, with a throughput of up to 250 slides per hour. In the same run, the instrument can handle special stains, immunohistochemistry, *in situ* hybridization and haematoxylin-eosin slides. The i6000 can operate in bar-code mode if required, and provides sufficient flexibility for individual pathologists to customize staining results.

Data Management

NuGenesis Technologies/ Spotfire

www.nugenesis.com; www.spotfire.com

Integrated platform for dynamic analysis

The integration of the NuGenesis Scientific Data Management system with Spotfire's DecisionSite analytic application provides scientists with the ability to automatically capture, process and visualize data from disparate sources using an integrated program. Information from different types of instruments and software programs can be correlated and analysed using this platform, which combines a relational database repository with an analytics software package.

Biomek FX

Beckman Coulter www.beckmancoulter.com

Sophisticated assay development

Automated Assay Optimization (AAO) software for the Biomek 2000 lab automation workstation has been enhanced with sophis-



The i6000: automating traditional histopathology techniques.

ticated assay scheduling capabilities. AAO facilitates the development of robust assays for high-throughput screening, with a claimed reduction of development time from months to days. Beckman Coulter's SAMI optimizing scheduler software is now embedded in the AAO program, enabling the scheduling of time-critical steps in the experiments.

Sample storage racks

Micronic www.micronic.com

Robust choices

The 'Premium Range' of 96-well tube racks from Micronic can accommodate alphanumerically (A1-H12) labelled 1.4 ml deep-well tubes. The racks are robust enough to be autoclaved over 100 times. The company also makes the 96-position Robo-Rack, which has a closed wall design, allowing robotic grippers to reach the rack from any height. The

Sure-Shot contoured well top allows positive tube lead-in (A1 to H12); any tubes not perfectly delivered to the desired rack location will not cause stoppages, but are positively guided into the correct well location.

ACTEvap

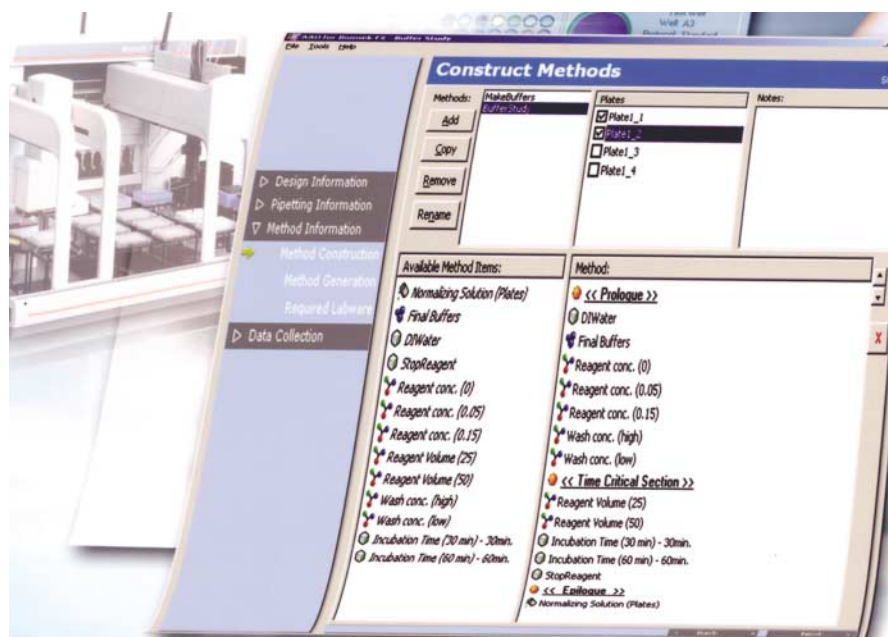
Advanced ChemTech

www.advancedchemtech.com

Vanishing act

ChemTech's ACTEvap unit can be used in conjunction with a rotary evaporator or a hot plate attached to a vacuum source to evaporate up to 33 samples simultaneously. The device comes with a series of interchangeable racks that will accommodate all common laboratory vial sizes (4, 20, 22 and 40 ml). The ACTEvap can be used to remove a wide range of solvents, including volatile liquids such as trifluoroacetic acid and very high boiling liquids, for example, demethylformamide.

These notes are compiled in the Nature office from information provided by the manufacturers. Press releases to be considered for publication can be sent to newproducts@nature.com.



Program listing: some of the AAO options.

The eternal molecule

As a prelude to the many celebrations around the world saluting the 50th anniversary of the discovery of the DNA double helix, *Nature* presents a collection of overviews that celebrate the historical, scientific and cultural impacts of a revelatory molecular structure.

Few molecules captivate like DNA. It enthral scientists, inspires artists, and challenges society.

It is, in every sense, a modern icon. A defining moment for DNA research was the discovery of its structure half a century ago. On 25 April 1953, in an article in *Nature*, James Watson and Francis Crick described the entwined embrace of two strands of deoxyribonucleic acid. In doing so, they provided the foundation for understanding molecular damage and repair, replication and inheritance of genetic material, and the diversity and evolution of species.

The broad influence of the double helix is reflected in this collection of articles. Experts from a diverse range of disciplines discuss the impact of the discovery on biology, culture, and applications ranging from medicine to nanotechnology. To help the reader fully appreciate how far the double helix has travelled, we also include the original landmark paper by Watson and Crick and the two accompanying papers by Maurice Wilkins, who shared the Nobel Prize with Watson and Crick in 1962, and by co-discoverer Rosalind Franklin, and their co-authors (pages 397–401). We are pleased to acknowledge the financial support of Roche in producing this collection of articles.

Transforming science

Given the immense significance of the double helix, it is difficult to imagine a world that wasn't transfixed by its discovery. Yet, as Robert Olby recalls on page 402, the proposed structure initially received a lukewarm reception. Maclyn McCarty, who, together with Oswald Avery and Colin MacLeod, had previously showed DNA to be the substance of inheritance, shares his personal perspective (page 406).

In science, where a lifetime's work can often be encapsulated in a few shining moments, the greatest controversies are sometimes over the sharing of credit. The discovery of the double helix is no exception. The premature death and posthumous treatment of Rosalind Franklin, whose X-ray images of DNA crystals revealed telltale clues of a double helical structure, propelled her portrayal as a feminist icon. But, as discussed here by her biographer Brenda Maddox (page 407), Franklin is better remembered as a committed and exacting scientist who saw no boundaries between everyday life and science.

Most of our readers will have grown up with the double helix, and yet it is still startling to consider how quickly DNA biology has progressed in just a lifetime. Bruce Alberts reviews how the elegant pairing of the two strands of the double helix revealed the mechanism for replicating the essential units of inheritance (page 431). Errol Friedberg considers the vulnerability of the DNA molecule to damage and the multitude of ways in which cells repair the damage (page 436). And

Gary Felsenfeld and Mark Groudine describe how the gargantuan DNA molecule is packaged inside the minuscule cells of the body, and how an additional layer of information is encrypted within the proteins intimately associated with DNA (page 448). It is perhaps salutary also to recognize what is still to be learnt about the physiological states in which DNA exists, as discussed by Philip Ball (page 421).

As reviewed by Leroy Hood and David Galas (page 444), DNA science generated the tools that spawned the biotechnology revolution. It enabled the cloning of individual genes, the sequencing of whole genomes and, with the application of computer science, transformed the nature and interactions of molecules into an information science. Carlos Bustamante and co-authors consider how we are still learning much about the distinct structural and physical properties of the molecule (page 423). And according to Nadrian Seeman, DNA may develop new applications as a material for nanoscale engineering (page 427).

Influencing society

Beyond scientific and technological forums, the double helix has imprinted on society's views of history, medicine and art. As discussed by Svante Pääbo (page 409), the records of evolution have been recalibrated with information traced through DNA sequence. On page 412, Aravinda Chakravarti and Peter Little revisit the 'nature versus nurture' debate and our developing view of the interplay between genetic and environmental factors in human disease. And DNA science will transform clinical medicine according to John Bell (page 414), providing a new taxonomy for human disease and triggering a change to health care practice. On page 440, Gustav Nossal reviews how an understanding of DNA processes, such as recombination, have transformed the field of immunology.

As a visual icon, and as a profound influence on our nature, the DNA molecule has permeated the imagery and art of our time, and is described by Martin Kemp (page 416) as the *Mona Lisa* of this scientific age. Given that broad impact, and revolutions that are yet to come, it is perhaps appropriate to leave the last word to an artist. Written in 1917, the poem *Heredity* by Thomas Hardy (see inset) seems to foreshadow both the essence and the fascination of the molecule that we celebrate here. □

Carina Dennis Commissioning Editor
Philip Campbell Editor, *Nature*

doi:10.1038/nature01396

Heredity

*I am the family face;
Flesh perishes, I live on,
Projecting trait and trace
Through time to times anon,
And leaping from place to place
Over oblivion.*

*The years-heired feature that can
In curve and voice and eye
Despise the human span
Of durance — that is I;
The eternal thing in man,
That heeds no call to die.*

Thomas Hardy
(First published in *Moments of Vision and Miscellaneous Verses*, Macmillan, 1917)

Quiet debut for the double helix

Robert Olby

Department of the History and Philosophy of Science, 1017 Cathedral of Learning, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA
(e-mail: olbyr+@pitt.edu)

Past discoveries usually become aggrandized in retrospect, especially at jubilee celebrations, and the double helix is no exception. The historical record reveals a muted response by the scientific community to the proposal of this structure in 1953. Indeed, it was only when the outlines appeared of a mechanism for DNA's involvement in protein synthesis that the biochemical community began to take a serious interest in the structure.

"... we may expect genetic chemistry to become in time an integrating core for cellular biochemistry."

Robert Sinsheimer, in a lecture delivered at the California Institute of Technology, 1956 (published in ref. 1, p. 1128).

To recall the year 1953 is to visit — and for some of us to revisit — another world, when *Nature* did not use the abbreviation DNA for deoxyribonucleic acid. In June that year, Elizabeth II, Queen of the United Kingdom, was crowned amidst much pomp and ceremony. In March, British scientists prepared to construct an atomic power station by the Calder River. Two months later, Mount Everest was conquered. At the University of London my biochemistry teacher enthused about Frederick Sanger's success in the first sequencing of the units of a protein, insulin. But deoxyribonucleic acid (DNA) was not even mentioned. Yet in 1953 *Nature* published seven papers on the structure and function of DNA²⁻⁸, but only one national British newspaper — the *News Chronicle* — referred to the double helix⁹ (see facsimile below).

Reception to the double helix

Fifty years on it is hard to believe the double helix had such a lukewarm reception. But turn to *Nature* and to *Science* in the 1950s and what do we find? Figure 1 records the number of papers in *Nature* reporting on any aspects of DNA, and of these the number that mention the Watson–Crick model or cite any of the 1953 papers on DNA structure. Through the decade *Nature's* volumes increased in size, and in 1960 the number of volumes published per year was doubled. This increase was accompanied by an increase in the number of papers on some aspect of DNA, but references to the double helix did not increase. The pattern of citation in *Science* is similar.

At the time the structure of DNA was discovered, there was already a considerable ongoing programme

of research on DNA (see time line in Box 1). These studies include the physical properties of DNA, methods of extraction, and whether the content and composition of DNA is the same for all the cells of the same organism. Also discussed were the damaging effects of ultraviolet light and ionizing radiation on DNA, and differing views over the involvement of nucleic acids in protein synthesis.

Researchers working on DNA at that time were principally biochemists and physical chemists, and their institutional locations and funding were chiefly medically related. Their interests and means of support related to two main concerns of the time — the action of 'mutagens' (agents that cause mutations in DNA), a subject important to the international debate on the effects of ionizing radiation and radioactive materials (see article in this issue by Friedberg, page 436), and the nature of protein synthesis, of great interest to biochemists in the light of its importance in growth and nutrition, in addition to cancer research.

In the light of the muted reception of the structure, let us take a different angle and ask what justification was there in the 1950s for giving the DNA double helix more than passing attention? At the time, most scientists reading *Nature* viewed DNA as a 'conjugated protein', owing to its association with protein; it was important as such, but not in its own right. This was despite the remarkable work of Oswald Avery, Colin MacLeod and Maclyn McCarty in 1944 (ref. 10; and see article in this issue by McCarty, page 406), followed by Al Hershey and Martha Chase's demonstration in 1952 (ref. 11) that most of the material entering a bacterium from an infecting bacterial virus is nucleic acid not protein. These studies made DNA look very much like the hereditary material.

Connecting structure to function

More information was needed to convince the scientific community. What was there about the chemistry of DNA to justify its role in inheritance? An answer came with the structure put forward by Watson and Crick. Chief among its "novel features" of "considerable biological interest"², Watson and Crick described the pairing of the bases, where adenine forms hydrogen bonds with thymine, and guanine with cytosine. This pairing, they wrote, "immediately suggests a possible copying mechanism for the genetic material."² Expanding on this in a subsequent paper appearing in *Nature* a month later, they wrote of DNA: "Until now, however, no evidence has been



Ritchie Calder's report on the discovery of the structure of DNA on page 1 of the *News Chronicle*, 15 May 1953.

No one suggests these groupings can yet be arranged artificially. Discovering how these chemical "cards" are shuffled and paired will keep the scientists busy for the next 50 years.



presented to show how it might carry out the essential operation required of a genetic material, that of exact self-duplication.”⁵

With these words Watson and Crick claimed their priority on a mechanism for DNA replication, but admitted there were problems with their scheme: how do the chains unwind and separate “without everything getting tangled”⁵? What is the exact mechanism by which gene duplication occurs? How does the genetic material “exert a highly specific influence on the cell”¹² when the sequence of bases assumed to encode the specificity is on the inside of the helical molecule?

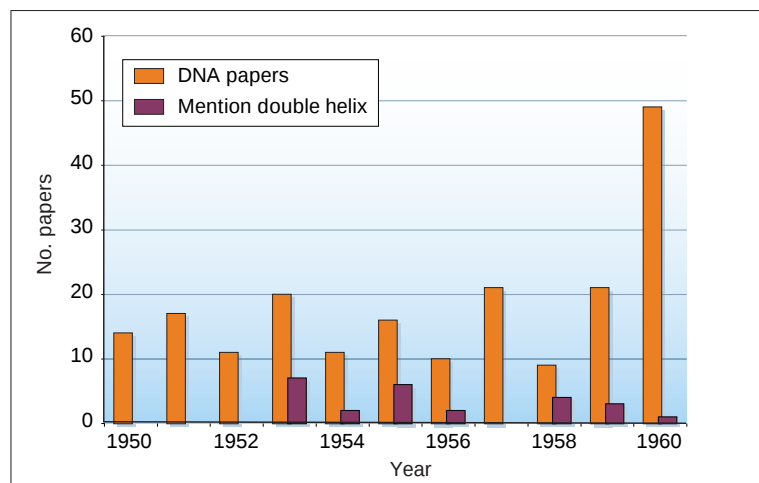
The ‘unwinding problem’ dominated much of the early discussions that followed the discovery of the DNA structure. In 1953, Watson and Crick admitted it was “formidable”¹², but support for their structure came in 1958, when Matthew Meselson and Franklin Stahl proved the semi-conservative nature of DNA replication¹³: each of the two new daughter DNA molecules formed during DNA replication consists of one strand from the original parent molecule and a new strand synthesized from the parent strand, which served as a template. This confirmed Watson and Crick’s theoretical prediction from the structure that replication would proceed in a semi-conservative manner. Later that same year, Arthur Kornberg announced the partial purification of an enzyme that catalyses DNA synthesis later called DNA polymerase¹⁴. This first linked enzymology to the double helix, for not long thereafter Kornberg provided biochemical evidence that DNA polymerase synthesizes new strands from opposite directions of the two chains of the molecule¹⁵.

In 1957, Crick defined biological ‘information’ as the sequence of the bases in the nucleic acids and of the amino acids in proteins, and proposed the now

famous ‘central dogma’ according to which information so defined flows between the nucleic acids and proteins only in one direction — from the former to the latter¹⁶. Just four years later, Marshall Nirenberg and Heinrich Matthaei successfully synthesized a polypeptide constituted of only one kind of amino acid (phenylalanine) using an RNA composed only of one kind of base (uracil). They concluded that “one or more [of these RNA bases] appear to be the code for phenylalanine.”¹⁷ Meanwhile, Crick, Sydney Brenner and Leslie Barnett had been using genetic analysis to investigate mutagenesis. This led them to the important concept of a form of mutation in which there is a ‘frame shift’ in the sequence of the bases in DNA, from which they went on to infer that the genetic message is composed of single or multiple triplets of bases, and that the message is read starting at a fixed point and proceeds always in the same

Six of the Nobel winners of 1962 display their diplomas after formal ceremonies in Stockholm’s Concert Hall. From left to right: Maurice Wilkins (Medicine), Max Perutz (Chemistry), Francis Crick (Medicine), John Steinbeck (Literature), James Watson (Medicine) and John Kendrew (Chemistry).

Figure 1 Papers published in *Nature* referring to DNA and the extent of their reference to the double helix 1950–1960.



Box 1

Time line of the discovery of the structure of DNA

- 1869 Fritz Miescher discovers that the nuclei of pus cells contain an acidic substance to which he gave the name 'nuclein'. Later he finds that nuclein is composed of a protein and a compound to which the name nucleic acid, and subsequently DNA, will be given.
- 1919 Phoebus Aaron Levene proposes the 'tetranucleotide' structure of DNA, whereby the four bases of DNA were arranged one after another in a set of four.
- 1928 Frederick Griffith finds that a substance in heat-killed bacteria can cause heritable changes in the live bacteria alongside them. He calls the phenomenon 'transformation'.
- 1938 Rudolf Signer, Torbjorn Caspersson and Einer Hammarsten find molecular weights for DNA between 500,000 and 1,000,000 daltons. Levene's tetranucleotide must be a polytetranucleotide.
- 1944 Oswald Avery, Colin MacLeod and Maclyn McCarty establish the chemical identity of Griffith's transforming principle as DNA, and they suggest that it may function as the genetic material.
- 1949 Erwin Chargaff reports that DNA base composition varies from one species to another, yet the ratio between the quantities of the two purine bases, adenine and thymine, and that between the quantities of the two pyrimidine bases, guanine and cytosine, remains about the same, namely one to one.
- 1949 Roger and Colette Vendrely, together with André Boivin find half as much DNA in the nuclei of sex cells as they find in the body cells, thus paralleling the reduction in the number of chromosomes, making DNA look like the genetic material.
- 1951 Rosalind Franklin distinguishes two forms of DNA, the paracrystalline B form and the crystalline A form.
- 1952 Al Hershey and Martha Chase find that DNA but scarcely any protein from an infecting bacterial virus enters the bacterial cell and can be recovered from the progeny virus particles.
- 1952 Rosalind Franklin and Raymond Gosling produce a magnificent X-ray diffraction pattern of the B form of DNA.
- 1953 James Watson and Francis Crick, Rosalind Franklin and Raymond Gosling, Maurice Wilkins, W. E. Seeds, Alec Stokes and Herbert Wilson, and Bertil Jacobson all publish on the structure of DNA²⁻⁸.
- 1954 George Gamow suggests a DNA code for the synthesis of proteins.
- 1955 Seymour Benzer analyses the fine structure of the genetic material of a bacterial virus at a level close to the distances that separate the individual bases along the DNA chain.
- 1957 Francis Crick proposes 'the sequence hypothesis' and 'the central dogma'.
- 1958 Matthew Meselson and Franklin Stahl demonstrate the semi-conservative replication of DNA.
- 1959 Arthur Kornberg and colleagues isolate the enzyme DNA polymerase.
- 1961 Marshall Nirenberg and Johann Heinrich Matthaei show that a sequence of nucleotide can encode a particular amino acid, laying the foundations for deciphering the genetic code.
- 1962 The Nobel prize in medicine is awarded to James Watson, Francis Crick and Maurice Wilkins.

direction¹⁸. Thus was the stage set for the subsequent unravelling of the entire genetic code.

From a muted reception in 1953 to accelerating momentum towards the end of the decade, one is tempted to infer that the DNA double helix was not taken seriously until a mechanism for its involvement in protein synthesis began to take shape. There was, to be sure, a small band of scientists who from the start either built their careers upon the implications of the structure (such as Meselson and Alexander Rich) or redirected their research to follow it up (including Seymour Benzer and Sydney Brenner). However, many scientists, notably Erwin Chargaff and

Alexander Dounce, did not refer to the structure in their scientific papers in the mid-fifties, even though it was clearly relevant and presumably known to them. Such omissions suggest that some biochemists had their own agendas, and the double helix was not at first seen as an aid to their work.

Biochemists debate protein synthesis

Biochemists' reservations about the double helix stemmed in part from the fact that evidential support for it in 1953 was far from strong. Watson and Crick themselves admitted that it "could in no sense be considered proved", although it was "most promising"¹⁹. In part the biochemists' coolness owed much to the debates among them over the mechanism of protein synthesis. The paper by Peter Campbell and Thomas Work, published in *Nature* on 6 June 1953, portrayed this debate vividly. They identified two contrasting theories under discussion on how proteins are made: first, the peptide theory (also known as the multi-enzyme theory), where proteins are made by "stepwise coupling of many small peptide units"; and second, the template theory, involving "synthesis on templates, each template being specific for a single protein structure and probably identifiable as a gene."²⁰

The peptide model was, for a very long time, supported by many prominent biochemists, including Joseph Fruton. The conviction behind it was the power of enzymes to both synthesize and break down their substrates, with a high degree of specificity attributed to both actions. Synthesis was proposed to involve the formation of a succession of peptides, ultimately yielding the protein molecule, and enzymes synthesize only those peptide bonds that they also hydrolyse. But the problem with this theory was that, except for a very few special cases, the alleged peptides constituting the intermediaries in protein synthesis could neither be detected in the cell nor incorporated into the protein being synthesized. Amino acids, however, could be incorporated, indicating they were the building blocks of proteins.

The second model of protein synthesis, which assumed synthesis on a template, had been advocated by Dounce in 1952. He pictured polypeptide chains being laid down on RNA molecules, and the RNA sequence determining the sequence of amino acids incorporated (on a one-to-one basis). Thus, DNA in the nucleus would control the order of bases in the RNA²¹.

After weighing up the merits and difficulties of Dounce's scheme, Campbell and Work voiced their distaste for the genetic control of protein synthesis, remarking in 1953 that: "...the gene is essentially an abstract idea and it may be a mistake to try to clothe this idea in a coat of nucleic acid or protein...if we must have a gene it should have a negative rather than a positive function so far as protein synthesis is concerned."²⁰ Only three years later, however, Robert Sinsheimer concluded a lecture at the California Institute of Technology with the following words: "The gene, once a formal abstraction, has begun to condense, to assume form and structure and defined activity."²¹

But those three years were a scene of pronounced change. By January 1957, when Fruton revised the second edition of his widely used textbook *General Biochemistry*, his remarks on the peptide theory were

cautious and were followed by a discussion of the role of RNA on which, he noted, there have been “stimulating speculations about the role of nucleic acids as ‘templates’ in protein synthesis.”²² Earlier in the book he devoted a paragraph to the double helix, describing it as an ‘ingenious speculation’. The only diagram was of the base pair adenine–guanine, rather than the helical model of the structure.

Kornberg had shown in 1957 that DNA replication follows the rules of base pairing, whereby DNA polymerase adds a base to the newly synthesized strand that is complementary to the opposing base in the template strand (A is always opposite T, and C always opposite G). But his interest in the subject had not been stimulated by Watson and Crick’s discovery. Rather, in 1953 he was preoccupied with how coenzymes (non-protein compounds needed for enzyme activity) are synthesized from nucleotides. He was led to wonder how DNA and RNA might be made from thousands of nucleotides. “The significance of the double helix,” he recalled, “did not intrude” into his work until 1956, after he had shown that a “moderately purified fraction” of what he was later to call DNA polymerase “appeared to increase the size of a DNA chain.”^{23,24}

Conclusion

The two once enigmatic processes — DNA replication and protein synthesis — intersected ongoing research programmes in the physical, organic and biological chemistry of the early 1950s. After the discovery of the double helix, those grappling with the problem of replication found its molecular foundation in the structure of DNA, although it took more than two decades to deduce the intricate mechanism of its operation in the cell (see article in this issue by Alberts, page 431). Those working on protein synthesis found the source of its specificity lay in the base sequence of DNA.

But why celebrate this one discovery? Why not celebrate the golden jubilee of Max Perutz’s solution to the ‘phase problem’ for proteins in 1953, without which the subsequent discovery of the structure of myoglobin and haemoglobin would not have been possible? What about the year 2005 for celebrating the golden jubilee of Sanger’s determination of the complete amino-acid sequence of a protein? Undoubtedly, the double helix has remarkable iconic value that has contributed significantly to its public visibility, something that has not been achieved by any of the protein structures (see article in this issue by Kemp, page 416). There is, too, a degree of notoriety attaching to the manner of its discovery and the characters involved that has given spice to the story, as widely publicized by James Watson’s account of the discovery in *The Double Helix*, published in 1968 (ref. 25), and Brenda Maddox’s recent illuminating biography of Rosalind Franklin²⁶. But there is a centrality about DNA that relates to the centrality of heredity in general biology.

The silver and golden jubilees of the Queen’s accession to the throne have come and gone, nuclear power stations are no longer being built in the United Kingdom, and mountaineer after mountaineer has ascended Mount Everest without a fanfare of press reports. But DNA is very much in the news — whether it be as a tool for studying evolution, a forensic test for rape, a source of genetic information or a path to designer drugs. And what better emblem or mascot is there for molecular biology than the double helix, and its spartan yet elegant representation in the original paper² from the pen of Odile Crick, Francis’s wife, fifty years ago? □

doi:10.1038/nature01397

1. Sinsheimer, R. L. First steps toward a genetic chemistry. *Science* **125**, 1123–1128 (1957).
2. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
3. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738–740 (1953).
4. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740–741 (1953).
5. Watson, J. D. & Crick, F. H. C. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
6. Franklin, R. E. & Gosling, R. G. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature* **172**, 156–157 (1953).
7. Jacobson, B. Hydration structure of deoxyribonucleic acid and its physico-chemical properties. *Nature* **172**, 666–667 (1953).
8. Wilkins, M. H. F., Seeds, W. E., Stokes, A. R. & Wilson, H. R. Helical structure of crystalline deoxypentose nucleic acid. *Nature* **172**, 759–762 (1953).
9. Calder, R. Why you are you: nearer the secret of life. *News Chronicle* p. 1 (15 May 1953).
10. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp. Med.* **79**, 137–158 (1944).
11. Hershey, A. D. & Chase, M. Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* **36**, 39–56 (1952).
12. Watson, J. D. & Crick, F. H. C. The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.* **18**, 123–131 (1953).
13. Meselson, M. & Stahl, F. W. The replication of DNA in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **44**, 671–682 (1958).
14. Lehman, I. R., Bessmann, M. J., Simms, E. S. & Kornberg, A. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. *J. Biol. Chem.* **233**, 163–170 (1958).
15. Kornberg, A. Biological synthesis of deoxyribonucleic acid: an isolated enzyme catalyzes synthesis of this nucleic acid in response to directions from pre-existing DNA. *Science* **131**, 1503–1508 (1960).
16. Crick, F. H. C. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
17. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polynucleotides. *Proc. Natl Acad. Sci. USA* **47**, 1558–1602 (1961).
18. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
19. Crick, F. H. C. & Watson, J. D. The complementary structure of deoxyribonucleic acid. *Proc. R. Soc. Lond. A* **223**, 80–96 (1954).
20. Campbell, P. N. & Work, T. S. Biosynthesis of proteins. *Nature* **171**, 997–1001 (1953).
21. Dounce, A. Duplicating mechanisms for peptide chain and nucleic acid synthesis. *Enzymologia* **15**, 251–258 (1952).
22. Fruton, J. S. *General Biochemistry* 2nd edn (Wiley, New York 1958).
23. Kornberg, A. *For the Love of Enzymes. The Odyssey of a Biochemist* (Harvard Univ. Press, Cambridge, MA, 1989).
24. Kornberg, A. in *A Symposium on the Chemical Basis of Heredity* (eds McElroy, W. D. & Glass, B.) 605 (Johns Hopkins Press, Baltimore, 1957).
25. Watson, J. D. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (Atheneum, New York, 1968). [Norton Critical Edition (ed. Stent, G. S.) published by Norton, New York & London, 1980.]
26. Maddox, B. *Rosalind Franklin. The Dark Lady of DNA* (Harper Collins, London 2002).

Discovering genes are made of DNA

Maclyn McCarty

The Rockefeller University, 1230 York Avenue, New York 10021, USA (e-mail: mccartm@rockefeller.edu)

Maclyn McCarty is the sole surviving member of the team that made the remarkable discovery that DNA is the material of inheritance. This preceded by a decade the discovery of the structure of DNA itself. Here he shares his personal perspective of those times and the impact of the double helix.

Editor's note — For a long time, biologists thought that 'genes', the units of inheritance, were made up of protein. In 1944, in what was arguably the defining moment for nucleic acid research, Oswald Avery, Maclyn McCarty and Colin MacLeod, at Rockefeller Institute (now University) Hospital, New York, proved that DNA was the material of inheritance, the so-called stuff of life. They showed that the heritable property of virulence from one infectious strain of pneumococcus (the bacterial agent of pneumonia) could be transferred to a noninfectious bacterium with pure DNA¹. They further supported their conclusions by showing that this 'transforming' activity could be destroyed by the DNA-digesting enzyme DNAase^{2,3}.

This work first linked genetic information with DNA and provided the historical platform of modern genetics. Their discovery was greeted initially with scepticism, however, in part because many scientists believed that DNA was too simple a molecule to be the genetic material. And the fact that McCarty, Avery and MacLeod were not awarded the Nobel prize is an oversight that, to this day, still puzzles.

"The pivotal discovery of 20th-century biology."

Joshua Lederberg, Rockefeller University, 1994, referring to the discovery by McCarty, Avery and MacLeod.

At the time of our discovery and publication in 1944 (ref. 1) of the research showing that DNA is heritable, my personal view, which I shared with MacLeod, was that there was little doubt that genes are made of DNA, and that this would ultimately be accepted. I was not sure of the best approach to use in pursuing research on the subject, but suspected that clarification of the structure of DNA was necessary.

But this was not an area of research in which I had received any training. Additionally, I had planned to make my career in disease-oriented research, and knowledge of the gene did not seem likely to become applicable in this area for some years. Thus, when invited to lead my own laboratory in the Rockefeller Hospital, investigating streptococcal infection and the pathogenesis of rheumatic fever, I decided to leave Avery's laboratory for this new position in July 1946.

Rollin Hotchkiss joined Avery at this point, and together with Harriett Taylor (a recent PhD graduate in genetics who had joined the laboratory in 1945), carried out studies increasing the purity of the transforming DNA mixture by further reducing any contaminating traces of protein. Together with other investigators, they also showed that properties of the pneumococcus other than just specific polysaccharide components of its cell wall could be transferred by the DNA preparations, indicating that the purified DNA also contained other genes of the bacterium.

Our findings continued to receive little acceptance for a variety of reasons, the most significant being that the work on the composition of DNA, dating back to its first identification 75 years earlier, had concluded that DNA was too limited in diversity to carry genetic information. Even those biologists who had considered the possibility had dropped the idea, and the prevailing dogma was that if genes are composed of a known substance, it must be protein.

There were a few biologists who took a different view, the most notable being Erwin Chargaff, who changed his area of research to DNA after reading our 1944 paper¹. His work revealed the great diversity in DNA

isolated from various sources, and that despite this diversity the amount of adenine always equalled that of thymine, and the amount of guanine that of cytosine. The latter finding was an important factor in the next significant advance in the field — the Watson–Crick determination of the double helical structure of DNA.

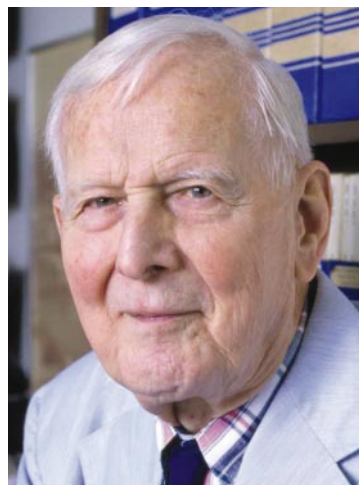
After the change in my research activity, I continued to give talks on our work on pneumococcal transformation and found the acceptance of the probable genetic role of DNA still to be minimal. However, I was convinced that it was only a matter of time before our results would become established.

Even though I was no longer involved in research on the subject, I continued to follow the developments as they appeared in the literature. Thus, when the papers of Watson and Crick describing the double helical structure of DNA were published in *Nature* in 1953, I certainly grasped the significance of their findings and was pleased to see such illuminating results come from a structural approach. I was not so pleased, however, that they failed to cite our work as one reason for pursuing the structure of DNA.

The concept of the double helix also hastened the silencing of those who had clung to the idea of genes as proteins. As a progressively larger body of investigators joined the study of the genetic role of DNA, there was an expanding amount of new information, starting with the resolution of the genetic code. By the end of the twentieth century, subsequent work on the mechanisms by which DNA is replicated with each cell division, is reshuffled with each generation, and is repaired when mistakes arise — the importance of which can in each case be traced back to the finding that DNA is the hereditary material — has transformed research in all areas of biology, technology and medicine. □

doi:10.1038/nature01398

1. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp. Med.* **79**, 137–158 (1944).
2. McCarty, M. Purification and properties of desoxyribonuclease isolated from beef pancreas. *J. Gen. Physiol.* **29**, 123–139 (1946).
3. McCarty, M. & Avery, O. T. Studies of the chemical nature of the substance inducing transformation of pneumococcal types II. Effect of desoxyribonuclease on the biological activity of the transforming substance. *J. Exp. Med.* **83**, 89–96 (1946).



Maclyn McCarty at The Rockefeller University.

The double helix and the 'wronged heroine'

Brenda Maddox

9 Pitt Street, London W8 4NX, UK (e-mail: bmaddox@pitt.demon.co.uk)

In 1962, James Watson, Francis Crick and Maurice Wilkins received the Nobel prize for the discovery of the structure of DNA. Notably absent from the podium was Rosalind Franklin, whose X-ray photographs of DNA contributed directly to the discovery of the double helix. Franklin's premature death, combined with misogynist treatment by the male scientific establishment, cast her as a feminist icon. This myth overshadowed her intellectual strength and independence both as a scientist and as an individual.

"Science and everyday life cannot and should not be separated. Science, for me, gives a partial explanation of life. In so far as it goes, it is based on fact, experience and experiment." Rosalind Franklin, in a letter to her father, summer 1940.

In late February 1953, Rosalind Franklin, a 33-year-old physical chemist working in the biophysics unit of King's College in London, wrote in her notebooks that the structure of DNA had two chains. She had already worked out that the molecule had its phosphate groups on the outside and that DNA existed in two forms.

Two weeks later James Watson and Francis Crick, at the Cavendish Laboratory at Cambridge, built their now celebrated model of DNA as a double helix. They did it not only through brilliant intuition and a meeting of compatible minds, but also on the basis of Franklin's unpublished experimental evidence, which had reached them through irregular routes. She did not know that they had seen either her X-ray photograph (Fig. 1), showing unmistakable evidence of a helical structure, or her precise measurements of the unit cell (the smallest repeating unit) of the DNA crystal.

As Watson was to write candidly, "Rosy, of course, did not directly give us her data. For that matter, no one at King's realized they were in our hands." When this admission appeared in Watson's best-selling, much-acclaimed book of the discovery, *The Double Helix*, published in 1968 (ref. 1), he was a Harvard professor and Nobel laureate (he had shared the prize for medicine and physiology in 1962, with Crick and Maurice Wilkins of King's College.) By then Franklin had died — in 1958, at the age of 37, from ovarian cancer.

Other comments dismissive of "Rosy" in Watson's book caught the attention of the emerging women's movement in the late 1960s. "Clearly Rosy had to go or be put in her place [...] Unfortunately Maurice could not see any decent way to give Rosy the boot". And, "Certainly a bad way to go out into the foulness of a [...] November night was to be told by a woman to refrain from venturing an opinion about a subject for which you were not trained."

A feminist icon

Such flamboyantly chauvinist phrases were sufficient to launch the legend of Franklin, the wronged heroine. So too was Watson's insistence on judging Franklin by her appearance rather than by her performance as a scientist. (She was, when she came to King's from the French government laboratory where she had worked from 1947 to the end of 1950, a recognized expert on



the structure of coals, carbons and disordered crystals, with many publications to her credit.)

The Franklin myth has continued to grow, abetted by the fact of her tragically early death. Franklin has become a feminist icon — the Sylvia Plath of molecular biology — seen as a genius whose gifts were sacrificed to the greater glory of the male. Her failure to win the Nobel prize has been given as a prime example

©1956-2003 THE NOVARTIS FOUNDATION (WWW.NOVARTISFOUNDATION.ORG.UK), FORMERLY THE CIBA FOUNDATION. REPRODUCED WITH PERMISSION.

Figure 1 “Her photographs are among the most beautiful X-ray photographs of any substance every taken.” — J. D. Bernal, 1958. Franklin’s X-ray diagram of the B form of sodium thymonucleate (DNA) fibres, published in *Nature* on 25 April 1953, shows “in striking manner the features characteristic of helical structures”⁵.

of the entrenched misogyny of the science establishment, rather than the consequence of the Nobel statute against posthumous awards.

Watson’s caricature of the bad-tempered “Rosy” drew a counter-blast from her good friend, the American writer Anne Sayre, in *Rosalind Franklin and DNA*, published in 1975 (ref. 2). Sayre’s book provided a much-needed corrective portrait, but was marred by a feminist bias. For example, it grossly underestimated the number of women scientists at King’s in the early 1950s. Sayre maintained there was only one other than Franklin, whereas there were at least eight on the senior staff. She insisted, moreover, that women’s exclusion from the King’s senior common room deprived Franklin of the intellectual companionship of her colleagues. In fact, most of the scientific staff preferred to eat in the joint dining room, men and women together, and the women, in general, felt well treated at King’s.

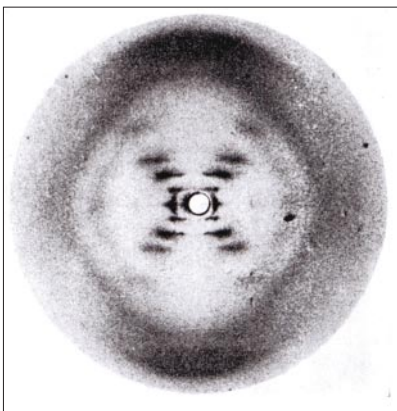
Reassessing the facts

As a biographer writing nearly three decades later and given access to Franklin’s personal correspondence, I found a more attractive, capable woman than Watson had suggested, and a King’s College more congenial and welcoming to women scientists than Sayre had allowed. I also found that Franklin felt singularly unhappy at King’s, not so much because of her gender, but because of her class and religion: a wealthy Anglo-Jew felt out of place in a Church of England setting dominated by swirling cassocks and students studying for the priesthood. “At King’s,” she wrote to Sayre (albeit inaccurately), “there are neither Jews nor foreigners”.

She was, in fact, so unhappy at King’s that, in early 1953, getting out as fast as possible was far more important to her than finishing her work on DNA. How far she had advanced was reported in two articles in *Nature*^{3,4} by Sir Aaron Klug, Franklin’s closest collaborator at Birkbeck College, London, where she moved to from King’s. He concluded that she had come very close to discovering the structure of DNA herself.

An irony of the story is that her own manuscript (coauthored by her student, R. G. Gosling and dated 17 March 1953) summarizing her results was already prepared by the time news reached King’s that Watson and Crick had cracked the DNA secret. Thus she inserted a hand-written amendment to her manuscript — which was published in *Nature* on 25 April 1953 (ref. 5), along with the now-celebrated Watson and Crick paper and another by Wilkins, Herbert Wilson and Alec Stokes of King’s — to say “Thus our general ideas are not inconsistent with the model proposed by Watson and Crick in the preceding communication”. And so they should have been, for the Watson-Crick findings were based on her data.

There is no evidence that she knew that in late January 1953 Wilkins had innocently shown her Photograph 51, with its stark cross of black reflections (Fig. 1), to Watson, who was visiting King’s. Nor did she know



that in February 1953 Max Perutz, then at the Cavendish Laboratory, had let Watson and Crick see his copy of the Medical Research Council’s report summarizing the work of all principal researchers, including Franklin’s.

At the same time there is no evidence that Franklin felt bitter about their achievement or had any sense of having been outrun in a race that nobody but Watson and Crick knew was a race. Indeed, she could accept the Watson-Crick model as a hypothesis only. She wrote in *Acta Crystallographica* in September 1953 that

“discrepancies prevent us from accepting it in detail”⁶.

Belated credit

Watson and Crick seem never to have told Franklin directly what they subsequently have said from public platforms long after her death — that they could not have discovered the double helix of DNA in the early months of 1953 without her work. This is all the more surprising in view of the close friendship that developed among the three of them — Watson, Crick and Franklin — during the remaining years of her life. During this time, she was far happier at non-sectarian Birkbeck than she ever was at King’s, and led a spirited team of researchers studying tobacco mosaic virus (TMV).

From 1954 until months before her death in April 1958, she, Watson and Crick corresponded, exchanged comments on each other’s work on TMV, and had much friendly contact. At Wood’s Hole, Massachusetts, in the summer of 1954 Watson offered Franklin a lift across the United States as he was driving to her destination, the California Institute of Technology. In the spring of 1956 she toured in Spain with Crick and his wife Odile and subsequently stayed with them in Cambridge when recuperating from her treatments for ovarian cancer. Characteristically, she was reticent about the nature of her illness. Crick told a friend who asked that he thought it was “something female”.

In the years after leaving King’s, Franklin published 17 papers, mainly on the structure of TMV (including four in *Nature*). She died proud of her world reputation in the research of coals, carbons and viruses. Given her determination to avoid fanciful speculation, she would never have imagined that she would be remembered as the unsung heroine of DNA. Nor could she have envisaged that King’s College London, where she spent the unhappiest two years of her professional career, would dedicate a building — the Franklin-Wilkins building — in honour of her and the colleague with whom she had been barely on speaking terms. □

doi:10.1038/nature01399

1. Watson, J. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (Atheneum, New York, 1968).
2. Sayre, A. *Rosalind Franklin and DNA* (W. W. Norton & Co., New York, 1975).
3. Klug, A. Rosalind Franklin and the discovery of the structure of DNA. *Nature* **219**, 808–810, 843–844 (1968).
4. Klug, A. Rosalind Franklin and the double helix. *Nature* **248**, 787–788 (1974).
5. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740–741 (1953).
6. Franklin, R. E. & Gosling, R. G. The structure of sodium thymonucleate fibres: I. The influence of water content. II. The cylindrically symmetrical Patterson function. *Acta Crystallogr.* **6**, 673–677, 678–685 (1953).

The mosaic that is our genome

Svante Pääbo

Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany
(e-mail: paabo@eva.mpg.de)

The discovery of the basis of genetic variation has opened inroads to understanding our history as a species. It has revealed the remarkable genetic similarity we share with other individuals as well as with our closest primate relatives. To understand what make us unique, both as individuals and as a species, we need to consider the genome as a mosaic of discrete segments, each with its own unique history and relatedness to different contemporary and ancestral individuals.

The discovery of the structure of DNA¹, and the realization that the chemical basis of mutations is changes in the nucleotide sequence of the DNA, meant that the history of a piece of DNA could be traced by studying variation in its nucleotide sequence found in different individuals and in different species. But it was not until rapid and inexpensive methods became available for probing DNA sequence variation in many individuals that the efficient study of molecular evolution in general — and of human evolution in particular — became feasible. Thus, the development in the 1980s of techniques for efficiently scoring polymorphisms with restriction enzymes and amplifying DNA^{2,3} enabled the study of molecular evolution to become a truly booming enterprise.

What follows is a personal and, by necessity, selective attempt to consider what the accelerating pace of exploration of human genetic variation over the past two decades has taught us about ourselves as a species, as well as some suggestions for what may be fruitful areas for future studies.

Primate relations

The first insight of fundamental importance for our understanding of our origins came from comparisons of DNA sequences between humans and the great apes. These analyses showed that the African apes, especially the chimpanzees and the bonobos, but also the gorillas, are more closely related to humans than are the orangutans in Asia⁴. Thus, from a genetic standpoint, humans are essentially African apes (Fig. 1). Although there had been hints of this from molecular comparisons of proteins^{5,6}, it was a marked shift from the earlier common belief that humans represented their own branch separate from the great apes.

Our sense of uniqueness as a species was further rocked by the revelation that human DNA sequences differ by, on average, only 1.2 per cent from those of the chimpanzees⁷, as a consequence of humans and apes sharing a recent common ancestry. It should be noted

that the dating of molecular divergences has uncertainties of unknown magnitude attached, not least because of calibration based on palaeontological data. Nevertheless, it seems clear that the human evolutionary lineage diverged from that of chimpanzees about 4–6 million years ago, from that of gorillas about 6–8 million years ago, and from that of the orangutans about 12–16 million years ago⁷. Before the advent of molecular data, the human–chimpanzee divergence was widely believed to be about 30 million years old.

In fact, we have recently come to realize that the relationship between humans and the African apes is so close as to be entangled. Although the majority of regions in our genome are most closely related to chimpanzees and bonobos, a non-trivial fraction is more closely related to gorillas⁷. In yet other regions, the apes are more closely related to each other than to us (Fig. 2). This is because the speciation events that separated these lineages occurred so closely in time that genetic variation in the first ancestral species, from which the gorilla lineage diverged, survived

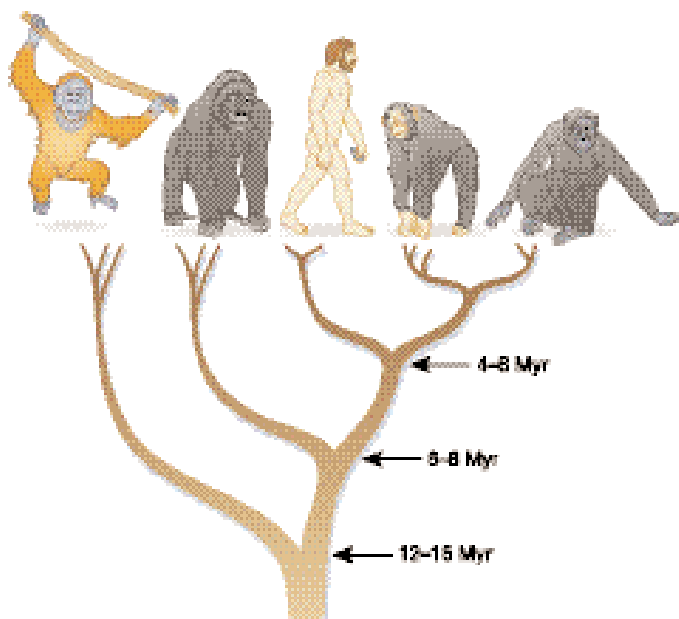
until the second speciation event between the human and chimpanzee lineages⁸. Thus, there is not one history with which we can describe the relationship of our genome to the genomes of the African apes, but instead different histories for different segments of our genome. In this respect, our genome is a mosaic, where each segment has its own relationship to that of the African apes.

Modern humans

The mosaic nature of our genome is even more striking when we consider differences in DNA sequence between currently living humans. Our genome sequences are about 99.9 per cent identical to each other. The variation found along a chromosome is structured in 'blocks' where the nucleotide substitutions are associated in so-called haplotypes (Figs 2b and 3). These 'haplotype blocks' are likely to result from the fact that recombination, that is, the re-shuffling of chromosome segments that occurs during formation of sex cells (meiosis), tends to occur in certain areas of the chromosomes more often than in others^{9–11}. In addition, the chance occurrence of recombination events at certain spots and not at others in the genealogy of human chromosomes will influence the structure of these blocks. Thus, any single human chromosome is a mosaic of different haplotype blocks, where each block has its own pattern of variation. Although the delineation of such blocks depends on the methods used to define them, they are typically 5,000–200,000 base pairs in length, and as few as four to five common haplotypes account for most of the variation in each block (Fig. 3).

Of 928 such haplotype blocks recently studied in humans from Africa, Asia and Europe¹², 51 per cent were found on all three

Figure 1 Tree showing the divergence of human and ape species. Approximate dates of divergences are given for, from left to right, orangutan, gorilla, human, bonobo and chimpanzee.



continents, 72 per cent in two continents and only 28 per cent on one continent. Of those haplotypes that were on one continent only, 90 per cent were found in Africa, and African DNA sequences differ on average more among themselves than they differ from Asian or European DNA sequences¹³. Therefore, within the human gene pool, most variation is found in Africa and what is seen outside Africa is a subset of the variation found within Africa.

Two parts of the human genome can be regarded as haplotype blocks where the history is particularly straightforward to reconstruct, as no recombination occurs at all. The first of these is the genome of the mitochondrion (the cellular organelle that produces energy and has its own genetic material), which is passed on to the next generation from the mother's side; the second is the Y chromosome, which is passed on from the father's side. Variation in DNA sequences from both the mitochondrial genome^{14–16} and the Y chromosome¹⁷, as well as many sections of the nuclear genome^{13,18–20}, have their geographical origin in Africa. Because other evidence suggest that humans expanded some 50,000 to 200,000 years ago²¹ from a population of about 10,000 individuals, this suggests that we expanded from a rather small African population. Thus, from a genomic perspective, we are all Africans, either living in Africa or in quite recent exile outside Africa.

Ancient humans

What happened to the other hominids that existed in the Old World from about 2 million years ago until about 30,000 years ago? For instance, the Neanderthals are abundant in the fossil record and persisted in western Europe until less than 30,000 years ago.

Analysis of Neanderthal mitochondrial DNA has shown that, at least with respect to the mitochondrial genome, there is no evidence that Neanderthals contributed to the gene pool of current humans^{22–25}. It is possible, however, that some as yet undetected interbreeding took place between modern humans and archaic hominids, such as *Homo erectus* in Asia or Neanderthals in Europe^{22,26,27}.

But any interbreeding would not have significantly changed our genome, as we know that the variation found in many haplotype blocks in the nuclear genome of contemporary humans is older than the divergence between Neanderthals and humans. Thus, the divergence of modern humans and Neanderthals was so recent that Neanderthal nuclear DNA sequences were probably more closely related to some current human DNA sequences than to other Neanderthals. In other words, the overlapping genetic variation that is likely to have existed between different ancient hominid forms makes it difficult to resolve the extent to which any interbreeding occurred.

Nevertheless, the limited variation among humans outside Africa, as well as palaeontological evidence²⁸, suggest that any contribution cannot have been particularly extensive. Thus, it seems most likely that modern humans replaced archaic humans without extensive interbreeding and that the past 30,000 years of human history are unique in that we lack the company of the closely related yet distinct hominids with which we used to share the planet.

Human variation and 'race'

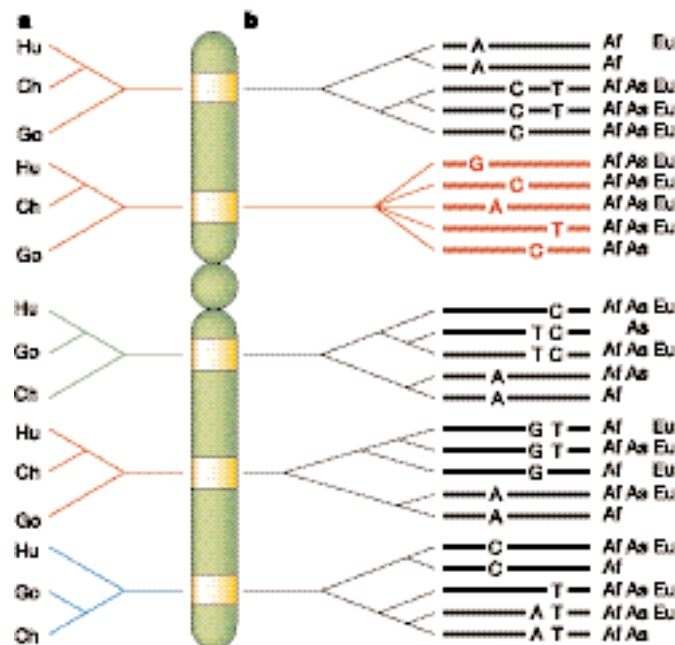
Comparisons of the within-species variation among humans and among the great apes have shown that humans have less genetic

variation than the great apes^{29,30}. Furthermore, early data that only about 10 per cent of the genetic variation in humans exist between so-called 'races'³¹ is borne out by DNA sequences which show that races are not characterized by fixed genetic differences. Rather, for any given haplotype block in the genome, a person from, for example, Europe is often more closely related to a person from Africa or from Asia than to another person from Europe that shares his or her complexion (for example, see ref. 32; Fig. 2).

Claims about fixed genetic differences between races (see ref. 33 for example) have proved to be due to insufficient sampling³⁴. Furthermore, because the main pattern of genetic variation across the globe is one of gene-frequency gradients³⁵, the contention that significant differences between races can be seen in frequencies of various genetic markers³⁶ is very likely due to sampling of populations separated by vast geographical distances. In this context it is worth noting that the colonization history of the United States has resulted in a sampling of the human population made up largely of people from western Europe, western Africa and southeast Asia. Thus, the fact that 'racial groups' in the United States differ in gene frequencies cannot be taken as evidence that such differences represent any true subdivision of the human gene pool on a worldwide scale.

Rather than thinking about 'populations', 'ethnicities' or 'races', a more constructive way to think about human genetic variation is to consider the genome of any particular individual as a mosaic of haplotype blocks. A rough calculation (Fig. 3) reveals that each individual carries in the order of 30 per cent of the entire haplotype variation of the human gene pool. Although not all of our

Figure 2 Within- and between-species variation along a single chromosome. **a**, The interspecies relationships of five chromosome regions to corresponding DNA sequences in a chimpanzee and a gorilla. Most regions show humans to be most closely related to chimpanzees (red) whereas a few regions show other relationships (green and blue). **b**, The among-human relationships of the same regions are illustrated schematically for five individual chromosomes. Most DNA variants are found in people from all three continents, namely Africa (Af), Asia (As) and Europe (Eu). But a few variants are found on only one continent, most of which are in Africa. Note that each human chromosome is a mosaic of different relationships. For example, a chromosome carried by a person of European descent may be most closely related to a chromosome from Asia in one of its regions, to a chromosome from Africa in another region, and to a chromosome from Europe in a third region. For one region (red), the extent of sequence variation within humans is low relative to what is observed between species. The relationship of this sequence among humans is illustrated as star-shaped owing to a high frequency of nucleotide variations that are unique to single chromosomes. Such regions may contain genes that contribute to traits that set humans apart from the apes.



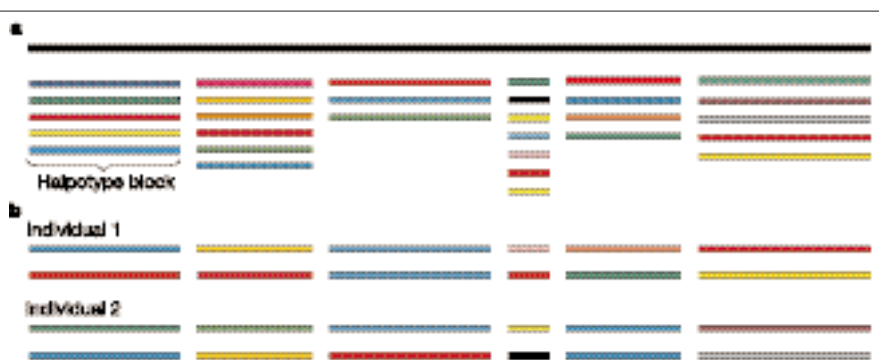


Figure 3 The mosaic structure of human genetic variation. **a**, Each human chromosome is made up of regions, called 'haplotype blocks', which are stretches of DNA sequence where three to seven variants (at frequencies above 5 per cent in the human population) account for most of the variation found among humans. Each such haplotype found in a block is illustrated here as a bar of different colour. The catalogue of haplotypes for every block makes up the 'haplotype map' of the human genome. **b**, The chromosomes of two hypothetical individuals are shown. Each individual carries two copies of each block (as humans carry two sets of chromosomes). As the chance that the two haplotypes carried at a block are identical is about 20 per cent, each of us carries an average of about 1.8 different haplotypes per block. Since there is on average 5.5 haplotypes for every block, each individual carries about 30 per cent of the total haplotype diversity of the entire human species. Haplotype blocks tend to be shorter in Africa than elsewhere; as a result, African variation will probably have to be used to define the species-wide block lengths, which may be an average of around 10,000 base pairs. Note that not all of the human genome may have a clearly definable haplotype-block structure.

genome may show a typical haplotype-block structure and more research is needed to fully understand the haplotype landscape of our genome, this perspective clearly indicates that each of us contain a vast proportion of the genetic variation found in our species. In the future, we therefore need to focus on individuals rather than populations when exploring genetic variation in our species.

Tracking human traits

What are the frontiers ahead of us in human evolutionary studies? One of them, to my mind, is to identify gene variants that have been selected and fixed in all humans during the past few hundred thousand years. These will include genes involved in phenotypic traits that set humans apart from the apes and at least some archaic human forms (for example, genes involved in complex cognitive abilities, language and longevity). However, an important obstacle in this respect is that there is little detailed knowledge of many of the relevant traits in the great apes. For example, only recently has the extent to which apes possess the capability for language³⁷ and culture³⁸ begun to be comprehensively described. As a consequence, we have come to realize that almost all features that set humans apart from apes may turn out to be differences in grade rather than absolute differences.

Many such differences are likely to be quantitative traits rather than single-gene traits. To have a chance to unravel the genetic basis of such traits, we will need to rigorously define the differences between apes and humans — for instance, how we learn, how we communicate and how we age. In the next

few years, geneticists will therefore need to consider insights from primatology and psychology, and more studies will be required that directly compare humans to apes.

There are, however, ways in which we can contribute towards the future unravelling of functionally important genetic differences between humans and apes. For example, we can identify regions of the human genome where the patterns of variation suggest the recent occurrence of a mutation that was positively selected and swept through the entire human population. The sequencing of the chimpanzee genome, as well as the haplotype-map project, will greatly help in this. Further prerequisites include the capability to determine the DNA sequence of many human genomes and the development of tools and methods to analyse the resulting data; in particular, a more realistic model of human demographic history is required.

Collectively these studies will allow us to identify regions in the human genome that have recently been acted upon by selection and thus are likely to contain genes contributing to human-specific traits (Fig. 2). Other interesting candidate genes for human-specific traits are genes duplicated or deleted in humans³⁹, genes that have changed their expression in humans⁴⁰, and genes responsible for disorders affecting traits unique to humans, such as language⁴¹ and a large brain size⁴².

A problem inherent in studying genes that are involved in traits unique to humans, such as language, is that functional experiments cannot be performed, as no animal model exists, and transgenic humans or chimpanzees cannot be constructed. A further difficulty is that many genes that enable

humans to perform tasks of interest may exert their effects during early development where our ability to study their expression both in apes and humans is extremely limited.

A challenge for the future is therefore to design ways around these difficulties. This will involve *in vitro* as well as *in silico* approaches that study how genes interact with each other to influence developmental and physiological systems. As these goals are achieved, we will be able to determine the order and approximate times of genetic changes during the emergence of modern humans that led to the traits that set us apart among animals. □

doi:10.1038/nature01400

1. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
3. Saiki, R. K. *et al.* Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
4. Miyamoto, M. M., Slightom, J. L. & Goodman, M. Phylogenetic relations of humans and African apes from DNA sequences in the *uvr*-globin region. *Science* **238**, 369–373 (1987).
5. Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA, 1963).
6. Wilson, A. C. & Sarich, V. M. A molecular time scale for human evolution. *Proc. Natl Acad. Sci. USA* **63**, 1088–1093 (1969).
7. Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**, 481–489 (2001).
8. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
9. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
10. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctuated meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
11. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
12. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
13. Yu, N. *et al.* Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).
14. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
15. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
16. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
17. Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
18. Stoneking, M. *et al.* Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061–1071 (1997).
19. Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
20. Takahata, N., Lee, S. H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
21. Harpending, H. & Rogers, A. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1**, 361–385 (2000).
22. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
23. Ovchinnikov, I. V. *et al.* Molecular analysis of Neandertal DNA from the northern Caucasus. *Nature* **404**, 490–493 (2000).
24. Krings, M., Geisler, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. DNA sequence of the mitochondrial hypervariable region II

- from the Neandertal type specimen. *Proc. Natl Acad. Sci. USA* **96**, 5581–5585 (1999).
25. Krings, M. *et al.* A view of Neandertal genetic diversity. *Nature Genet.* **26**, 144–146 (2000).
 26. Nordborg, M. On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63**, 1237–1240 (1998).
 27. Pääbo, S. Human evolution. *Trends Cell Biol.* **9**, M13–M16 (1999).
 28. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357**, 563–579 (2002).
 29. Deinard, A. & Kidd, K. Evolution of a HOXB6 intergenic region within the great apes and humans. *J. Hum. Evol.* **36**, 687–703 (1999).
 30. Kaessmann, H., Wiebe, V., Weiss, G. & Pääbo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet.* **27**, 155–156 (2001).
 31. Lewontin, R. C. The problem of genetic diversity. *Evol. Biol.* **6**, 381–398 (1972).
 32. Kaessmann, H., Heissig, F., von Haesler, A. & Pääbo, S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genet.* **22**, 78–81 (1999).
 33. Harris, E. E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
 34. Yua, N. & Li, W.-H. No fixed nucleotide difference between Africans and non-Africans at the pyruvate dehydrogenase E1 α -subunit locus. *Genetics* **155**, 1481–1483 (2000).
 35. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1993).
 36. Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biological research: genes, race and disease. *Genome Biol.* **3**, 2007.1–2007.12 (2002).
 37. Tomasello, M. & Call, J. *Primate Cognition* (Oxford Univ. Press, New York, 1997).
 38. Whiten, A. *et al.* Cultures in chimpanzees. *Nature* **399**, 682–685 (1999).
 39. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
 40. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
 41. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
 42. Jackson, A. P. *et al.* Identification of microcephalin, a protein implicated in determining the size of the human brain. *Am. J. Hum. Genet.* **71**, 136–142 (2002).

Acknowledgements

My work is funded by the Max Planck Society, the Bundesministerium für Bildung und Forschung and the Deutsche Forschungsgemeinschaft. I thank B. Cohen, H. Kaessmann, D. Serre, M. Stoneking, C. Stringer, L. Vigilant and especially D. Altshuler for helpful comments on the manuscript.

Nature, nurture and human disease

Aravinda Chakravarti* & Peter Little†

*McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 600 North Wolfe Street, Jefferson Street Building, 2-109, Baltimore, Maryland 21287, USA (e-mail: aravinda@jhmi.edu)

†School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia (e-mail: p.little@unsw.edu.au)

What has been learnt about individual human biology and common diseases 50 years on from the discovery of the structure of DNA? Unfortunately the double helix has not, so far, revealed as much as one would have hoped. The primary reason is an inability to determine how nurture fits into the DNA paradigm. We argue here that the environment exerts its influence at the DNA level and so will need to be understood before the underlying causal factors of common human diseases can be fully recognized.

"We used to think our fate was in our stars. Now we know, in large measure, our fate is in our genes." J. D. Watson, quoted in *Time* magazine, 20 March 1989 (ref. 1).

The double helix, in its simplicity and beauty, is the ultimate modern icon of contemporary biology and society. Its discovery provided the bridge between the classical breeding definition and the modern functional definition of genetics, and permanently united genetics with biochemistry, cell biology and physiology. The DNA structure provided an immediate explanation for mutation and variation, change, species diversity, evolution and inheritance. It did not, however, automatically provide a mechanism for understanding how the environment interacts at the genetic level.

One gene, one disease

Recognition that genes have a role in human disease dates back to the rediscovery of the rules that govern the inheritance of genes by Gregor Mendel — the so-called Mendelian

laws of inheritance. So far, human geneticists have been most successful at understanding single-gene disorders, as their biological basis, and thus presumed action, could be predicted from inheritance patterns. Mendelian diseases are typically caused by mutation of a single gene that results in an identifiable disease state, the inheritance of which can readily be traced through generations.

The landmark sequencing of the human genome provided some important lessons about the role of genes in human disease. Notably, mutations in specific genes lead to specific biological changes, and rarely do mutations in multiple genes lead to an identical set of characteristics that obey 'Mendelian inheritance'. Additionally, sequence diversity of mutations is large and, consequently, individual mutations are almost always rare, showing relatively uniform global distributions.

But a few exceptions do exist. Some recessive mutations (mutations that influence a person only if both copies of the gene are altered) are surprisingly common in specific populations. This defiance of general mutation patterns arises either from chance increases in frequency in isolated populations, such as in the Old Order Amish², or from the protective effect of a deleterious mutation in a single copy, such as the genetic mutation that on the one hand causes sickle-cell anaemia, but on the other hand offers protection against malaria³. These examples show that human history, geography and ecology of a particular people are relevant to understanding their present-day molecular disease burden⁴.

For over 90 years, the association between DNA mutations and a vast variety of single-gene disorders has repeatedly emphasized the notion that human disease results from faults in the DNA double helix (see, for example, the Online Mendelian Inheritance in Man database at www.ncbi.nlm.nih.gov/omim/, which provides a catalogue of human genes and genetic disorders). Is it then too extrapolative to suggest that all diseases and traits, each of which has some familial and imputed inherited component, will be caused by a corrupted piece of double helix?

Is our fate encoded in our DNA?

Is Watson's genetic aphorism of human disease really true? The excitement of genetics, and the perceived medical importance of the human genome sequence, is pegged to the promise of an understanding of common chronic disease and not rare Mendelian diseases. In theory, one might hope that approaches used successfully to identify single-gene diseases could simply be applied to the common causes of world-wide morbidity and mortality, such as cancer, heart disease, psychiatric illness and the like. This would enable a boon for diagnosis, understanding and the eventual treatment of these common maladies⁵.

The reality is that progress towards identifying common disease mutations has been slow, and only recently have there been some successes⁶. It is now appreciated that although genes are one contributor to the origin of common diseases, the mutations they contain must have properties that are different from the more familiar, deterministic features of single-gene mutations. Indeed, the underlying genes are likely to be numerous, with no single gene having a major role, and mutations within these genes being common and imparting small genetic effects (none of which are either necessary or sufficient⁷).

Moreover, there is a suspicion that these mutations both interact with one another and with the environment and lifestyle, although the molecular specificity of inter-

actions is unproven⁸. To complicate matters, common disorders frequently show large population differences that have led to health disparities and, as is becoming more evident, the incidence of these disorders can show significant changes over time⁹.

Interplay of DNA and environment

The inability of geneticists to easily identify common disease genes has been seen as a vindication of the importance of nurture. This is too simplistic; the influence of nature and nurture cannot be neatly divided, as it is clear that nurture is important to biology through its actions on DNA and its products. The environment must affect the regulation of critical genes by some mechanism and so, seen another way, mutations are not the only agent for altering gene function.

The scientific literature of cancer research reveals that despite having heterogeneous origins — both inherited and acquired — a specific tumour develops only from altering the expression (activity) of specific sets of genes¹⁰. That is, a variety of exposures and mutations collaborate to change the activity of specific genes and, consequently, interrupt precise aspects of cell metabolism. The regulation of circadian rhythm is another example of how external environmental cues influence DNA functions¹¹.

Thus, the double helix inevitably interacts with the environment, directly and indirectly, to predispose or protect us from disease. If perturbations of multiple genes contribute to a disorder, then the activities of these genes can be affected by any combination of mutation and environmental exposure altering their function. It is our opinion that genes have a stronger, maybe even a pervasive, role in all diseases and traits, with the understanding that it is the collective action of genes and nurture that underpins ultimate disease outcome.

Rather than dismissing the role of environment, our view embraces it directly, and, by that, expands the meaning of the term 'genetic'. It also emphasizes the work that remains to be done to understand gene regulation and, in particular, how genes and their products are modulated by external cues and how homeostasis is disrupted in human disease. Human beings are each the product of a unique genome and a unique set of experiences. Both need to be understood to intervene effectively in disease causation.

Implications for medicine

What does this mean in practice? The assessment of the quantitative role of genes in human traits is derived largely from studies on identical and fraternal twins (Fig. 1). By this measure, all common disorders have a 'genetic' basis, but the contribution varies from slight in some cancers and multiple sclerosis,

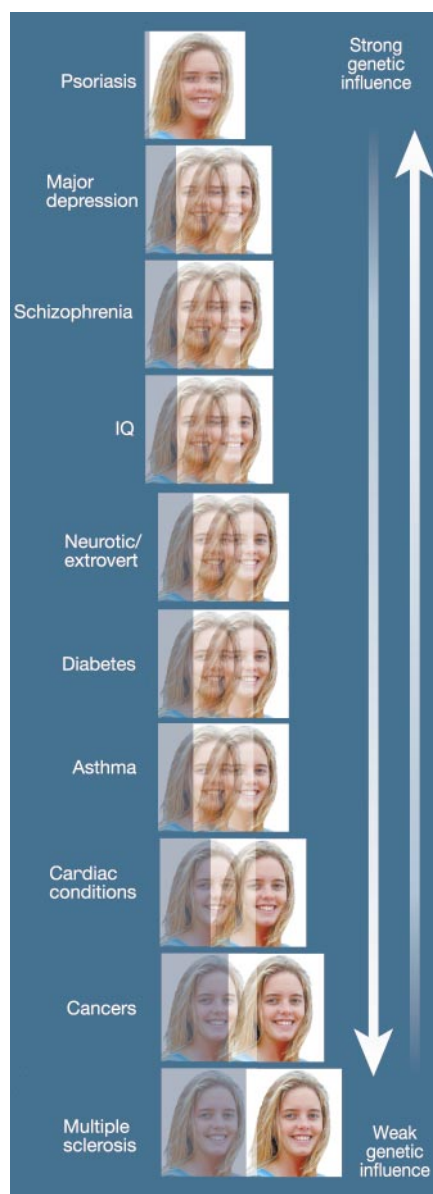


Figure 1 Studies of identical twins have revealed that some conditions, such as psoriasis, have a strong genetic component and are less influenced by environmental and lifestyle factors — identical twins are more likely to share these diseases. But other conditions, such as multiple sclerosis, are only weakly influenced by genetic makeup and therefore twins may show differences depending on their exposure to various environmental factors.

to moderate in diabetes, heart diseases, migraine and asthma, to high in disorders such as psoriasis¹². Critically, the discordance between identical twins — where twins show different diseases despite being genetically identical — illustrates the influence of exogenous factors, but does not prove the lack of influence of genes: of course, environmental factors over a lifetime affect an individual's chance of developing disease.

Let us assume, for the sake of argument, that all of the relevant genetic and environmental factors are identified that lead to a disease. Appreciating the relationship of

genetic variation and environment suggests that a number of presently fashionable ideas about genetics are simplistic; two in particular are the 'bar code' view of genetic diagnosis and the 'right medicine for the right patients'.

Common genetic variations are essentially binary — either an adenine or guanine base, or a cytosine or thymine base — at a given position in the sequence. Unfortunately, this leads to a tendency to define genetic individuality as a binary pattern, a so-called 'bar code' for each individual. Some genetic variants convey susceptibility to a disease, but they typically convey risk rather than certainty of being afflicted with a condition.

Knowledge based on the sequence could have significant public health implications, and even be predictive at the population level. But a human DNA bar code would provide uncomfortable, perhaps even intolerable, knowledge of likely outcomes, with no certainty, only probabilities. Most individuals, we suspect, are ill equipped to deal with the knowledge that they have a 50 per cent chance of succumbing to an illness; equally, society has had great difficulty in knowing how to respond to such information, hence the concerns regarding genetic discrimination¹³. The reality is that the genetic bar code is weakly predictive and individuals may find this threatening, life enhancing or just irrelevant; in any event, much work is needed to enable the predictive revolution in medicine.

Human genetic individuality has forced the recognition that medicine has to refocus on the individual. This has been the rallying cry, particularly within the pharmaceutical business, of pharmacogenomics (the application of genome-scale understanding to the development of medicines), and there is no doubt that understanding of the variation within drug-metabolizing enzymes has exploded in the past 20 years¹⁴. The underpinning idea is enormously attractive — if genetic analysis of key DNA variations can be used to understand how individuals might respond to drugs, then it could be possible to eliminate the difficult, sometimes lethal, hit-and-miss approaches to medication that are a necessary feature of present medical practice.

Unfortunately, the influence of lifestyle is just as much a feature of drug response as it is of any other genetically influenced condition. The classic case of the influence of drinking grapefruit juice on the levels of many drugs¹⁵ illustrated that there can be no such thing as 'the patient', because the patient is living in a complex world that changes by the minute. Once again, predictions for the population do not have the same predictive power for individuals.

Future challenges

The challenges that lifestyle presents to genetic studies are considerable. We believe

that the next 50 years will bring a genuine revolution of far greater individual significance than that delivered by genetics over the past 50 years. This is because lifestyle can conceivably be analysed, and in so doing, it should be possible to develop a genuinely personalized medicine.

Researchers can now think seriously about how to identify lifestyle influences: such studies will have to be on an unprecedented scale and one of the first of these, proposed to comprise 500,000 individuals in the United Kingdom, has already started¹⁶. These kinds of studies are a bold venture into relatively uncharted territory and face substantial technical, biological and science-culture challenges.

Scientifically, it is necessary to understand a deceptively simple equation: genes + environment = outcome. The difficulty here is the uncertainty surrounding both terms in the equation; ideally, one set of genetic factors will interact with one set of environmental influences to produce identical outcomes, but it is unknown whether this is always going to be the case. A far more difficult relationship would exist if multiple genetic factors interacted with multiple environments to achieve the same outcome. The example of glutathione S-transferase mutations, smoking and incidence of lung cancer¹⁷ shows it is possible to detect some interactions, but it is unclear how, or even if, statistical methods might be developed for addressing the more complex possibilities.

Perhaps the greatest unknown in undertaking these projects is human psychology; the consequences of smoking have been known for many decades, but people still smoke. Advice does not imply acceptance. How to turn knowledge into practical outcomes must be an increasing focus of attention for both researchers and funding agencies.

Psychology is also in play in the initial decision to undertake this research; for researchers, funding agencies and politicians there is great risk implicit in undertaking a hugely expensive project with complex outcome. People would like to live in a simpler world, with simpler decisions, but the vision of such a project is enormous: once complete, as much will be known about the origins of human disorders as can be discovered by using such epidemiological and genetic studies. Perhaps more important, the beginnings of a new medicine will emerge, one focused uniquely and completely upon the individual, upon the combination of genetic uniqueness and personal choices that are the very essence of individual lives.

If we are collectively bold in our present decisions and accept the risk of action, a world can be created where medicine is a guide, not a place of last resort. If the past 50 years has seen the revolution of DNA, then the revolution cannot be completed without

an appreciation of both genetic and environmental individuality; only then will individuals understand the meaning of their inheritance. □

doi:10.1038/nature01401

1. Jaroff, L. The gene hunt. *Time* 20 March, 62–67 (1989)
2. Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin. Genet.* **61**, 233–247 (2002).
3. Aidoo, M. *et al.* Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* **359**, 1311–1312 (2002).
4. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nature Rev. Genet.* **3**, 611–621 (2002).
5. Chakravarti, A. Single nucleotide polymorphisms: ...to a future of genetic medicine. *Nature* **409**, 822–823 (2001).
6. Carrasquillo, M. M. *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature Genet.* **32**, 237–244 (2002).
7. Cox, N. J. Challenges in identifying genetic variation affecting susceptibility to type 2 diabetes: examples from studies of the calpain-10 gene. *Hum. Mol. Genet.* **10**, 2301–2305 (2001).

8. Sullivan, P. F. *et al.* Analysis of epistasis in linked regions in the Irish study of high-density schizophrenia families. *Am. J. Med. Genet.* **105**, 266–270 (2001).
9. Boyle, J. P. *et al.* Projection of diabetes burden through 2050: impact of changing demography and disease prevalence in the United States. *Diabetes Care* **24**, 1936–1940 (2001).
10. Mürger, K. Disruption of oncogene/tumor suppressor networks during human carcinogenesis. *Cancer Invest.* **20**, 71–81 (2002).
11. Panda, S., Hogenesch, J. B. & Kay, S. A. Circadian rhythms from flies to human. *Nature* **417**, 329–335 (2002).
12. MacGregor, A. J. *et al.* Twins: novel uses to study complex traits and genetic diseases. *Trends Genet.* **16**, 131–134 (2000).
13. Wertz, D. C. Ethics watch. *Nature Rev. Genet.* **3**, 496 (2002).
14. Evans, W. E. & Johnson, J. A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genet.* **2**, 9–39 (2001).
15. Lown, K. S. *et al.* Grapefruit juice increases felodipine oral availability in humans by decreasing intestinal CYP3A protein expression. *J. Clin. Invest.* **99**, 2545–2553 (1997).
16. Wright, A. F., Carothers, A. D. & Campbell, H. Gene-environment interactions—the BioBank UK study. *Pharmacogenomics J.* **2**, 75–82 (2002).
17. Stucker, I. *et al.* Genetic polymorphisms of glutathione S-transferases as modulators of lung cancer susceptibility. *Carcinogenesis* **23**, 1475–1481 (2002).

The double helix in clinical practice

John I. Bell

*The Office of the Regius Professor of Medicine, University of Oxford, Oxford OX3 9DU, UK
(e-mail: regius@medsci.ox.ac.uk)*

The discovery of the double helix half a century ago has so far been slow to affect medical practice, but significant transformations are likely over the next 50 years. Changes to the way medicine is practised and new doctors are trained will be required before potential benefits are realized.

"It is much more important to know what kind of patient has a disease than to know what kind of disease a patient has." Caleb Parry, 18th century physician, Bath.

The structure of DNA established the basic framework that would develop into the field of molecular genetics. The information gleaned from this scientific endeavour continues to have a profound influence on our understanding of biological systems¹. As most human diseases have a significant heritable component, it was soon recognized that the characterization of the genetic determinants of disease would provide remarkable opportunities for clinical medicine, potentially altering the way disease was understood, diagnosed and treated.

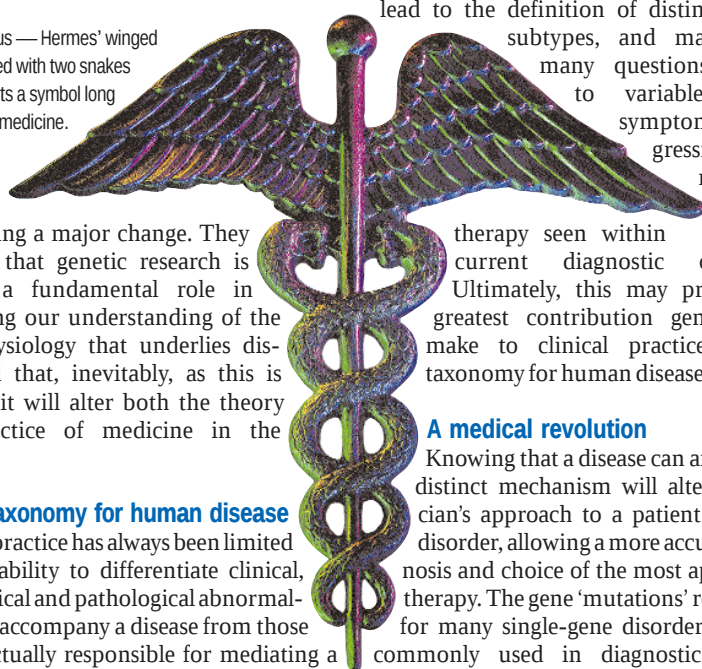
But despite the obvious potential applications to medicine, the development of significant genetic advances relevant to clinical practice could take generations. This is in marked contrast to many other medical-related discoveries that occurred around the same time and which were translated rapidly into clinical practice. For instance, the development of penicillin by Ernst Chain and Howard Florey in 1941 was saving thousands of lives within months of their discovery of how to efficiently produce the antibiotic². Discoveries relating to disease aetiology, such as the recognition in 1950 of a relationship between smoking and lung cancer, have had a profound effect on mortality³.

This was despite the convictions of at least one distinguished statistical geneticist who argued against the causality of this observation, implying that a common genetic factor caused both lung cancer and a predilection to smoking cigarettes⁴!

Although other important discoveries have had demonstrably more impact on health care at the time of their fiftieth anniversaries than has the double helix, its slower transition from discovery to clinical implementation will be balanced by its potentially profound impact across all medical disciplines. Progress has been slow, but mounting evidence suggests that, while public health and antibiotics produced important healthcare outcomes in the past 50 years, the next 50 are likely to belong to genetics and molecular medicine.

The potential impact of genetics on clinical practice has been questioned by some observers⁵ who believe that the positive predictive value of genetic testing for most common disease genes will be insufficient to provide the beneficial effects seen with single-gene disorders, which affect only a tiny proportion of the population. Many advocates of genetics argue, on the other hand, that our understanding of disease is

The caduceus — Hermes' winged staff entwined with two snakes — represents a symbol long adopted by medicine.



undergoing a major change. They contend that genetic research is playing a fundamental role in improving our understanding of the pathophysiology that underlies disease and that, inevitably, as this is applied, it will alter both the theory and practice of medicine in the future⁶.

A new taxonomy for human disease

Clinical practice has always been limited by its inability to differentiate clinical, biochemical and pathological abnormalities that accompany a disease from those events actually responsible for mediating a disease process. Clinicians may have moved on from calling 'fever' a disease⁷, but they still rely on phenotypic criteria to define most diseases, and yet these may obscure the underlying mechanisms and often mask significant heterogeneity. As Thomas Lewis pointed out in 1944, diagnosis of most human disease provides only "insecure and temporary conceptions"⁸. Of the main common diseases, only the infectious diseases have a truly mechanism-based nomenclature.

An understanding of the genetic basis of maladies is providing a new taxonomy of disease, free from the risk that the diagnostic criteria related to events are secondary to the disease process, rather than to its cause. Genetic information has allowed us to identify mechanistically distinct forms of diabetes, defining an autoimmune form of the disease associated with human leukocyte antigens (a highly diverse complex of immune-system genes), and recently has implicated dysfunction of factors that affect both expression and modification of gene products in mediating the adult form of the disorder⁹. Similarly, we are now aware of a range of molecules and pathways previously not recognized in the pathogenesis of asthma^{10–12}.

A clearer understanding of the mechanisms and pathways that mediate disease will

lead to the definition of distinct disease subtypes, and may resolve many questions relating to variable disease symptoms, progression and response to

therapy seen within current diagnostic categories. Ultimately, this may provide the greatest contribution genetics will make to clinical practice: a new taxonomy for human disease.

A medical revolution

Knowing that a disease can arise from a distinct mechanism will alter a physician's approach to a patient with that disorder, allowing a more accurate prognosis and choice of the most appropriate therapy. The gene 'mutations' responsible for many single-gene disorders are now commonly used in diagnostic practice, whereas those associated with common complex diseases are just being characterized. Although their predictive value will be less than with single-gene disorders, their contribution as risk factors will be similar to other risk factors such as blood pressure, cholesterol levels and environmental exposures. Because much of clinical practice involves evaluating and acting on risk probabilities, the addition of genetic risk factors to this process will be an important extension of existing practice. The overall effect of genetic risk factors is likely to be significant. For example, recent estimates in breast cancer suggest that the attributable genetic risks are likely to exceed the predictive value of a range of existing non-genetic risk factors¹³.

Other potential applications of genetics in health care may be realized in a shorter timeframe. Individual variation in response to drugs and in drug toxicity is a significant problem, both in clinical practice and in the development of new therapeutic agents. Clear examples now exist of genetic variants that alter metabolism, drug response or risk of toxicity^{14,15}. Such information provides an opportunity to direct therapy at individuals most likely to benefit from an intervention, thereby reducing cost and toxicity, and improving methods for drug development.

The discovery of the structure of DNA not only led to an ability to characterize genetic determinants in disease, but also provided the tools necessary for the revolution in molecular medicine that has occurred in the past 25 years. The description of the double helix was the first important step in the development of techniques to cut, ligate and amplify DNA. The application of these molecular biology and DNA-cloning techniques has already had a profound impact on our understanding of the basic cellular and molecular processes that underlie disease.

Molecular biology has improved our ability to study proteins and pathways involved in disease and has provided the technology necessary to generate new sets of targets for small-molecule drug design. It has also enabled the creation and production of a new range of biological therapeutics — recombinant proteins such as interferon, erythropoietin and insulin, as well as therapeutic antibodies, which are one of the fastest growing classes of new treatments. Further extensions of this methodology will see the inevitable introduction of DNA-based therapies that will produce proteins of interest in the appropriate cellular setting. DNA-based vaccines represent the first wave of such novel gene-therapy approaches to disease and many more are expected to follow.

We are undergoing a revolution in clinical practice that depends upon a better understanding of disease mechanisms and pathways at a molecular level. Much has already been achieved: an enhanced understanding of disease-related pathways, new therapies, novel approaches to diagnostics and new tools for identifying those at risk. But more remains to be done before the full impact of genetics on medicine is realized. Complex disease, with multiple susceptibility determinants (both environmental and genetic), will take time to dissect. This information must then be moved into the clinic and evaluated for its benefits.

As the practice of medicine moves to one more scientifically founded in disease mechanisms, many aspects of clinical practice will need to be transformed. Individual genetic variation is likely to explain a significant part of the heterogeneity seen clinically in the natural history of disease and in response to therapy. Tools to tailor medicine to an individual's needs rather than directing it at a population will inevitably become available. Similarly, as predictions of risk improve, early or preventative therapy of high-risk populations will become a reality, with screening programmes targeted to those at particularly high risk.

Transforming clinical practice

For fundamental changes to take place in clinical practice, sweeping transformation will be needed to healthcare provision, economic management and training. It is currently difficult to predict the cost-benefit ratio for such changes — certainly the present impact of molecular medicine has not made medicine less expensive. Few medical schools adequately train their students to think mechanistically about disease; indeed, the trend towards pattern-recognition medicine, away from basic science training, means that we are still far from educating the next generation of clinicians to apply the knowledge and tools bequeathed to us by the double helix. The evolution in health care that will incorporate these new principles of early diagnosis and individualized therapy will be a daunting

Table 1 Molecular genetics in clinical practice

• Mechanistically based diagnostic criteria
• Predisposition testing and screening
• Rapid molecular diagnostic testing of pathogens
• Pharmacogenetics
• Identification of new drug targets
• Tools for molecular medicine (for example, recombinant DNA methodology)
• Recombinant expression of therapeutic proteins
• Gene therapy

challenge in an era of uncertainty for healthcare systems worldwide.

The influence of genetic and molecular medicine on the health of patients is already sufficiently ubiquitous that it will have an impact on most common diseases. Its influence will grow over the next few decades (Table 1). It will not, however, answer all of the questions about human health, nor will it provide all the answers for optimizing clinical practice. The reductionism that accompanies molecular genetics will identify the pieces in the jigsaw, but assembling these to understand how complex systems malfunction will require a substantially more integrated approach than is available at present.

The crucial role played by environmental determinants of disease will perhaps become more tractable when combined with an understanding of genetic susceptibility. Sceptics, rightly, will wish to see more data before they acknowledge that molecular medicine will be truly transformed over the next 50 years, despite the fact that its influence on diagnostics and new therapeutics is already clearly apparent. A transition is underway, the direction of travel is clear, but managing the change in clinical practice may

prove at least as challenging as resolving the original structure of the helix. □

doi:10.1038/nature01402

1. Watson, J. D. & Crick, F. C. H. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Abraham, E. P. et al. Further observations on penicillin. *Lancet* **ii**, 177–188 (1941).
3. Doll, R. & Hill, A. B. Smoking and carcinoma of the lung. *Br. Med. J.* **2**, 739–748 (1950).
4. Fisher, R. A. Cancer and smoking. *Nature* **182**, 596 (1958).
5. Holtzman, N. A. & Marteau, T. M. Will genetics revolutionise medicine? *N. Engl. J. Med.* **343**, 141–144 (2000).
6. Bell, J. I. The new genetics in clinical practice. *Br. Med. J.* **316**, 618–620 (1998).
7. Osler, W. *The Principles and Practice of Medicine* (Appleton, New York, 1892).
8. Lewis, T. Reflections upon medical education. *Lancet* **i**, 619–621 (1944).
9. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
10. Van Eerdekewegh, P. et al. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418**, 426–430 (2002).
11. Cookson, W. O. C. M., Sharp, P. A., Faux, J. A. & Hopkin, J. M. Linkage between immunoglobulin E responses underlying asthma and rhinitis and chromosome 11q. *Lancet* **1**, 1292–1295 (1989).
12. Shirakawa, I. et al. Atopy and asthma: genetic variants of IL-4 and IL-13 signalling. *Immunol. Today* **21**, 61–64 (2000).
13. Pharoah, P. D. P. et al. Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genet.* **31**, 33–36 (2002).
14. Splawski, I. et al. Variant of SCN5A sodium channel implicated in risk of cardiac arrhythmia. *Science* **297**, 1333–1336 (2002).
15. Weber, W. W. *Pharmacogenetics* (Oxford Univ. Press, New York, 1997).



The *Mona Lisa* of modern science

Martin Kemp

Department of the History of Art, University of Oxford, Littlegate House, St Ebbes, Oxford OX1 1PT, and Wallace Kemp, Artakt, Studio D, 413–419, Harrow Road, London W9 3QJ, UK (e-mail: martin.kemp@trinity.ox.ac.uk)

No molecule in the history of science has reached the iconic status of the double helix of DNA. Its image has been imprinted on all aspects of society, from science, art, music, cinema, architecture and advertising. This review of the *Mona Lisa* of science examines the evolution of its form at the hands of both science and art.

"A monkey is a machine that preserves genes up trees, a fish is a machine that preserves genes in water; there is even a small worm that preserves genes in German beer mats. DNA works in mysterious ways." Richard Dawkins in *The Selfish Gene* (Oxford University Press, 1976).

History has thrown up a few super-images, which have so insinuated themselves into our visual consciousness that they have utterly transcended their original context. This is epitomized by the *Mona Lisa*, painted by Leonardo da Vinci around 1503. The double helix of DNA is unchallenged as the image epitomizing the biological sciences. Both images speak to audiences far beyond their respective specialist worlds, and both carry a vast baggage of associations.

In the worlds of popular image diffusion, particularly on the Internet, the double helix is beginning to rival the *Mona Lisa* as a playground for eccentrics and obsessives (Fig. 1). There is an apparent difference, of course. Leonardo's panel painting is the

product of human artifice, whereas DNA is a naturally occurring, large organic molecule. But Leonardo claimed that his art represented a systematic remaking of nature on the basis of a rational understanding of causes and effects. His painting is the result of a complex, nonlinear interaction between concept, subject, plan of action, acquired knowledge, skill, medium and the evolving image itself. In *The Art of Genes*¹, Enrico Coen argues that "biological development and human creativity are highly interactive processes in which events unfold rather than being necessarily pre-planned or anticipated. In other words, in both cases there is no easy separation between plan (or programme) and execution."

Looking at the investigation and representations of the double helix, we can say that

they are cultural activities no less than any painting. Behind the discovery lies the vast infrastructure of a scientific culture that led to the development of the knowledge, theories, institutions, techniques and equipment that made the quest both possible and desirable. The very natures of scientific models and representations, using whatever technique, are integral to the vehicles of science communication. Their visual look is compounded from a complex set of factors, ranging from technical to aesthetic. But, in case anyone should be getting the wrong impression, I acknowledge that the cultural vehicles are designed to deliver non-arbitrary information that is open to rational scrutiny as a way of working towards real knowledge of the physical constitution of the world.

Looked at from a popular perspective (and even from the standpoint of reputation within science), James Watson and Francis Crick are identified with DNA no less than Leonardo is identified with the *Mona Lisa*. The researchers were in a very real sense the 'authors' or 'artists' of the acts of visualization that generated their models of the molecule. But their brilliant achievement was not necessarily of a higher order than that of the other pioneers of molecular modelling, such as the Braggs, John Kendrew, Max Perutz, Maurice Wilkins and Linus Pauling. Rather, they were uniquely fortunate that their molecule was both visually compelling, as a supreme example of nature's 'sculpture', and

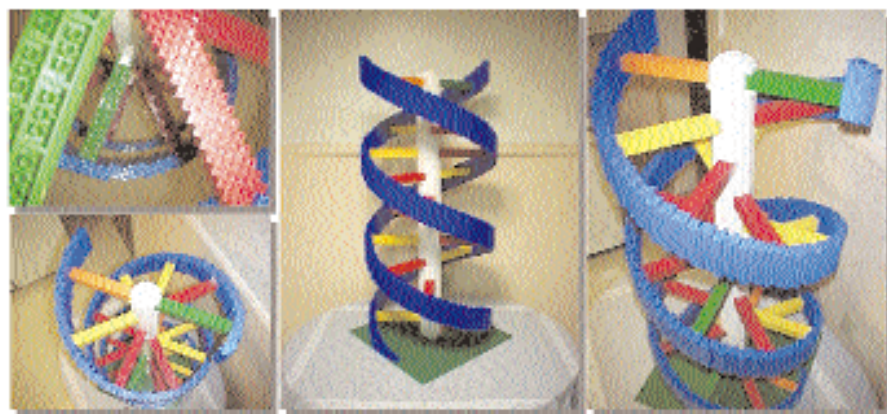


Figure 1 LEGO model of the DNA double helix (in reverse!) by Eric Harshbarger (2001), who also used his mastery of the coloured units of LEGO to compose a 'pixelated' LEGO version of the *Mona Lisa*. (Images courtesy of E. Harshbarger.)



Figure 2 Structure of DNA, drawn by Francis Crick's wife Odile Crick, which was published as the sole figure in Watson and Crick's seminal paper in *Nature*, 25 April 1953 (ref. 2).

lay at the heart of the twentieth-century version of the quest to unravel the ultimate secret of life.

The 50-year journey of the DNA molecule from the reticent line diagram in Watson and Crick's seminal article² (Fig. 2) to its position in today's world of global imagery is extraordinary. It is therefore timely to look at some of the representational issues involved in science communication, and then at a few selected instances of the various guises in which the molecule has replicated itself within varied visual habitats.

A model of communication

Looking back on the laconic article in *Nature* that announced the structure of DNA, which we tend to assume in retrospect provided the definitive solution, it is remarkable how little was actually given away. This is true of the article's sole diagram, drawn by Odile Crick, Francis's wife, which represented the sugar chains as directional ribbons, while the bases were rudimentary rods represented flat on (Fig. 2). Along the vertical axis runs the central pole, depicted as a thick line that is broken where the bases lie in front. This axis is a visually useful point of reference, but its early ubiquity seems to depend on the structural necessities of physical models. The developed model, composed from standard brass components with tailor-made metal bases, provided a more detailed and explicit entity for debate and large-scale publicity, although the famous photographs by Anthony Barrington Brown (Fig. 3), taken for an article in *Time* magazine, were actually staged a few months later.

The model of the double helix — like those of other molecules, such as the model of haemoglobin by Perutz — played an important role in scientific understanding, being both based upon and in turn affecting the acts of scientific conceptualization. Overtaken by more refined models made at King's College London, including the widely illustrated space-filling model with Van der Waals surfaces by Wilkins (Fig. 4), the ramshackle masterpiece of Watson and Crick

passed the way of so many obsolete bits of scientific paraphernalia. When, 23 years after its making, some of the specially cut plates (Fig. 5) resurfaced in Bristol, they were incorporated into a pious reconstruction by Farooq Hussain of King's College. Like an ancient Greek vase reassembled from chards,

the semi-original model is now a treasured cultural icon, displayed in the Science Museum in London.

Communicating the complex structure and, in due course, the awesomely intricate behaviour of the modular molecule, has provided an unparalleled challenge for

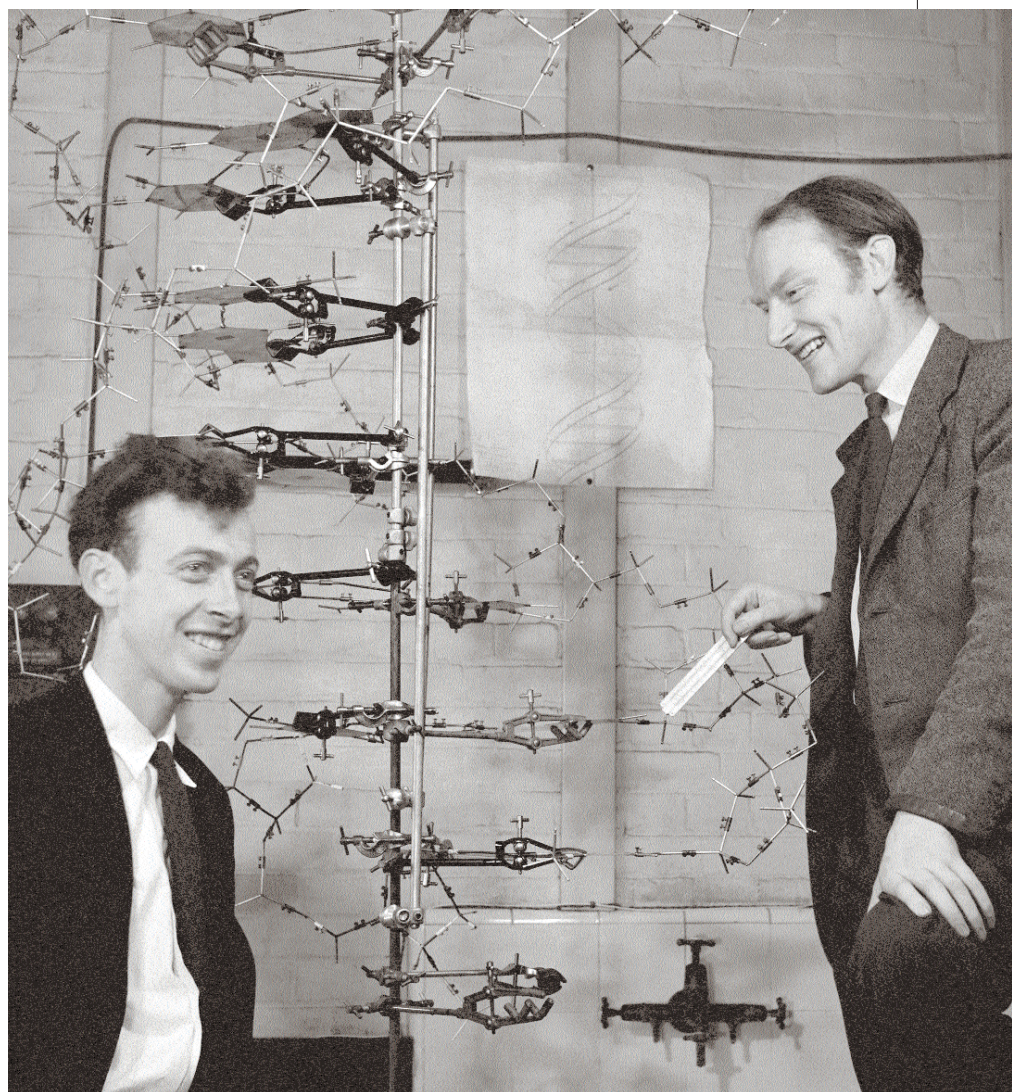


Figure 3 Anthony Barrington Brown's photograph of Watson and Crick with their model of DNA at the Cavendish Laboratory in Cambridge, 21 May 1953.

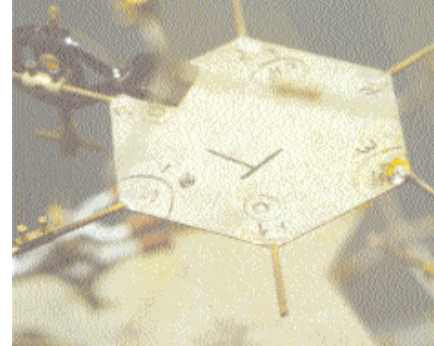


Figure 5 One of the specially cut plates used by Watson and Crick in their model of the structure of DNA.



Figure 6 Cover of *Nature* human genome issue, published on 15 February 2001. The image, by Eric Lander, was created by Runaway Technology Inc. using PhotoMosaic by Robert Silvers from original artwork by Darryl Leja (courtesy of the Whitehead Institute for Biomedical Research). Gregor Mendel, James Watson and Francis Crick are amongst the crowd.

Similarly, any hint of the double twist in any logo of a laboratory or biotech company is immediately identifiable.

Aesthetics and meaning

Given the role of aesthetic intuitions in the processes that led to its discovery, and its recognition as 'right', it is understandable that the double helix has itself assumed the guise of a work of art, not least in three-dimensional form. For artists, the attraction of a form that is both beautiful and full of all kinds of scientific and social significance is considerable.

Some grand structures have been commissioned by academic institutions, whereby the artist has basically been given a brief to make a sculpture representing DNA, much like a sculptor might be commissioned to produce an anatomically accurate image of a naked figure. For example, in 1998 Roger Berry produced a huge sculpture hanging down the central well of a multi-story staircase at the University of California at Davis (Fig. 7). Another rendering of the helical structure, *Spirals Time — Time Spirals*, resides on a hillock in the grounds of the Cold Spring Harbor Laboratory (Fig. 8). Designed by the artist, architect and theorist of post-

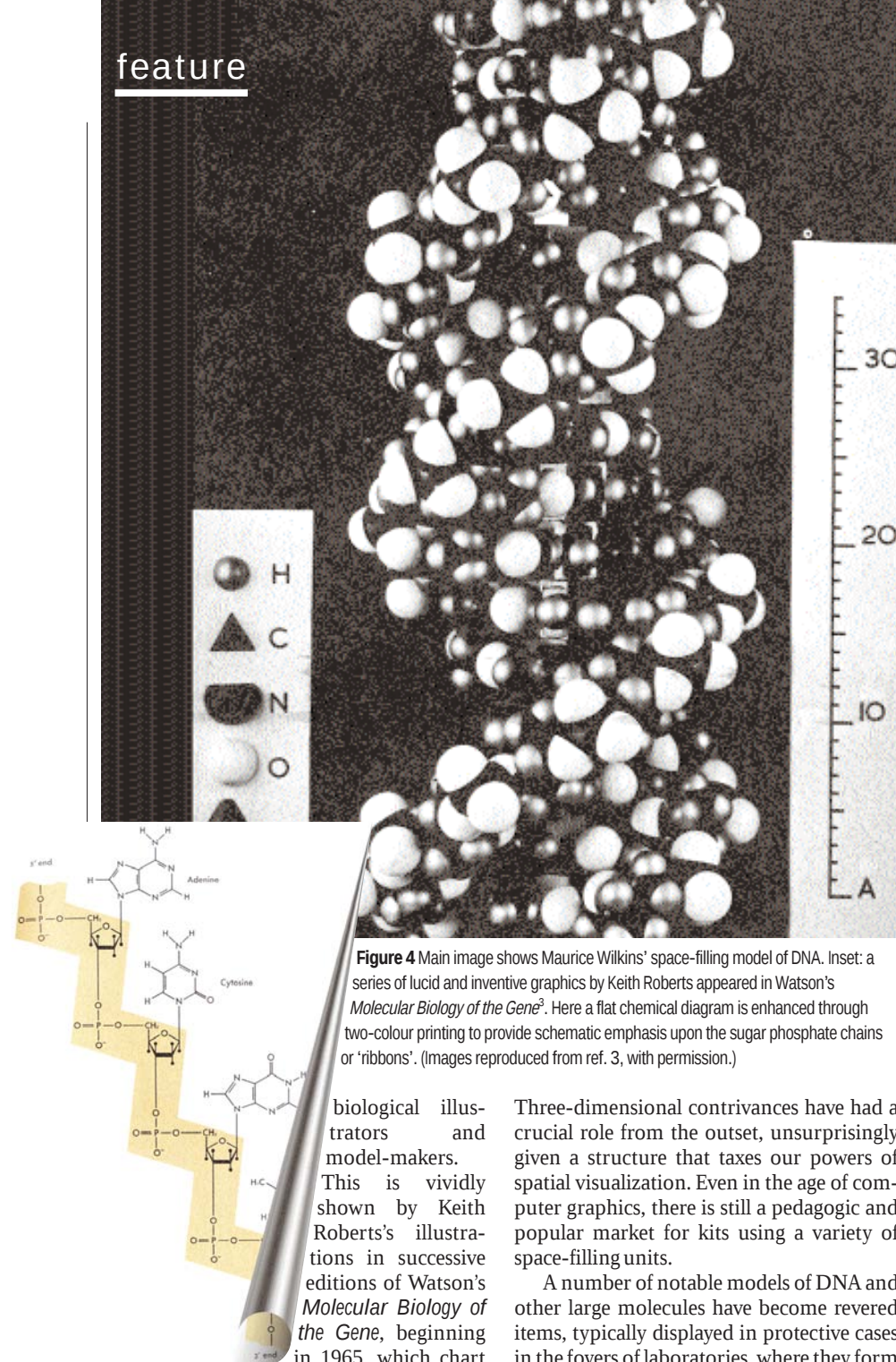


Figure 4 Main image shows Maurice Wilkins' space-filling model of DNA. Inset: a series of lucid and inventive graphics by Keith Roberts appeared in Watson's *Molecular Biology of the Gene*³. Here a flat chemical diagram is enhanced through two-colour printing to provide schematic emphasis upon the sugar phosphate chains or 'ribbons'. (Images reproduced from ref. 3, with permission.)

biological illustrators and model-makers.

This is vividly shown by Keith Roberts's illustrations in successive editions of Watson's *Molecular Biology of the Gene*, beginning in 1965, which chart

the complex interplay between developing science, graphic ingenuity and technologies of reproduction³ (see Fig. 4, inset).

As the complex functioning of DNA became increasingly elucidated, so methods and conventions of illustration that privileged behaviour over structure played an ever more conspicuous role. As with sets of illustrations in any science, the visual conventions not only reflect what scientists want to show, but also provide an important framework for thinking and visualization in the process of research itself. Subsequently, the resources of computer design, stereoscopy and, in particular, animation have provided a vivid sense of spatial and temporal processes, only partly possible in conventional text and illustration.

Three-dimensional contrivances have had a crucial role from the outset, unsurprisingly given a structure that taxes our powers of spatial visualization. Even in the age of computer graphics, there is still a pedagogic and popular market for kits using a variety of space-filling units.

A number of notable models of DNA and other large molecules have become revered items, typically displayed in protective cases in the foyers of laboratories, where they form part of the visual furniture that speaks of the enterprise of biological science in general and that of the institution in particular. For the sub-species of biologist known as 'molecular', the seductive geometry of DNA helps to underline the fundamental 'hardness' of their science, compared to natural historians and ecologists from whom they have become institutionally distinguished. It is in this spirit, less of didactic instruction than of emblematic signalling, that the double helix has become the icon for the communication of a generalized message. Few have any trouble in recognizing the ghostly twist that emerges from the mosaic of faces on the cover of the *Nature* issue devoted to the human genome on 15 February 2001 (Fig. 6).

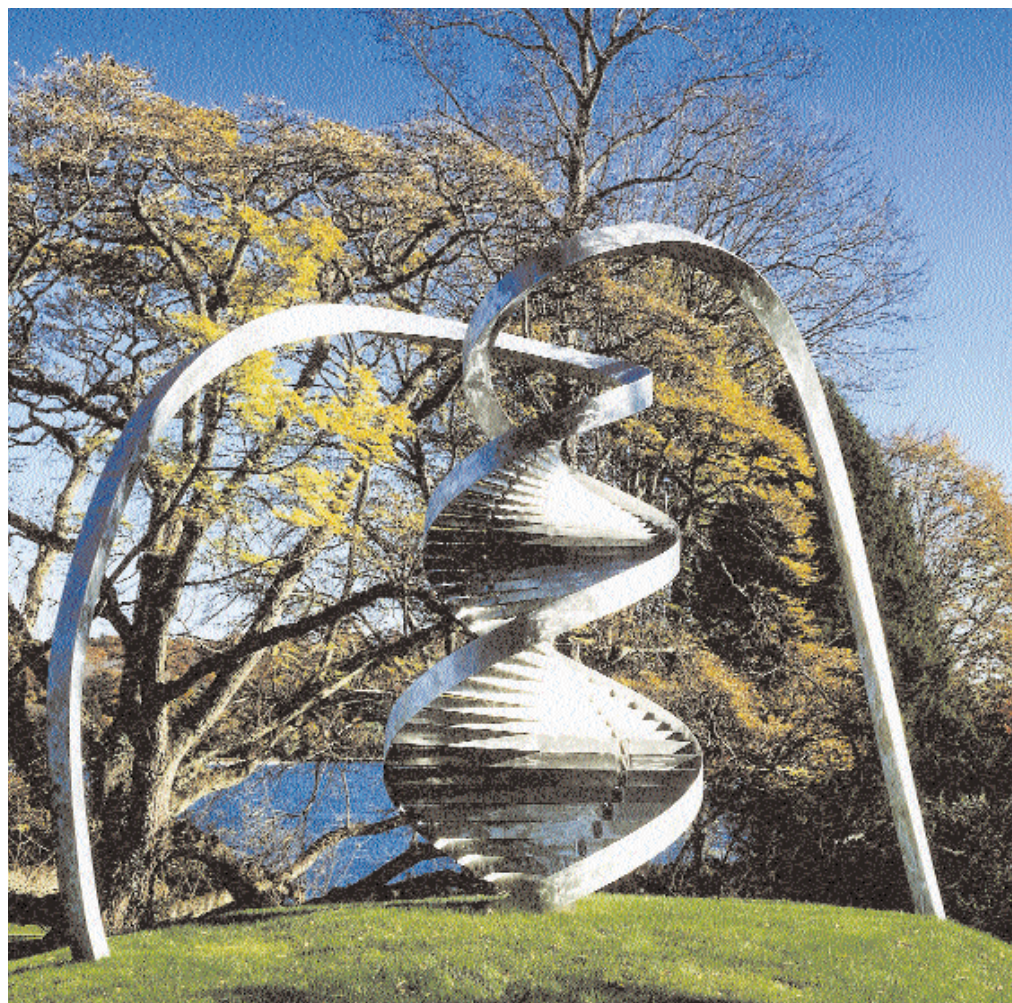


Figure 7 (left) *Portrait of a DNA Sequence* by Roger Berry (1998) at the Life Sciences Addition building, University of California, Davis. **Figure 8** *Spirals Time — Time Spirals* by Charles Jencks (2000) at Cold Spring Harbor Laboratory.

PERRY A. JOHNSON/UC DAVIS

modernism, Charles Jencks, it stands at the heart of a programme of commissioning and collecting artwork that expresses the vision of Watson — who became director of Cold Spring Harbor Laboratory in 1968 and president in 1994 — of an environment in which the visual stimulation of the surroundings is integral to the conduct of high-level mental activity.

In pursuit of structural aesthetics, the British sculptor, Mark Curtis, proposed a reformed molecular structure for DNA. As an artist concerned with geometrical logic and symmetries, Curtis was worried about the 'ugly' engineering of the Watson–Crick version. Rather than using the sugar phosphate backbones to control the helices, he proposed that stacked base pairs, coupled in an opposite orientation from the accepted bonding, comprises a helix of pentagonal plates around a central void of decagonal cross-section. The geometrical and structural probity of Curtis's models, which eschew a central pole, made it on to a British millennium stamp (Fig. 9), if not into the world of scientific orthodoxy. In a real sense, the molecular biologists' rejection of Curtis's effort to re-design DNA on the basis of a *a priori* principles represents an

extreme example of the tension within science itself between the polar instincts of the modelers and the empiricists.

Alongside such sculptural exploitations of the inherent beauty of the double helix has run a strand of artistic iconography that has been more overtly concerned with meaning. The tone for the more fantastical exploitations was set by the flamboyant surrealist, Salvador Dali, as ever concerned with the metaphysical potential implicit in scientific imagery. During the late 1950s and 1960s, the DNA molecule features as a symbolic vision, lurking in a surreal hinterland between galactic mystery and spiritual significance (as a kind of Jacob's Ladder). His *Butterfly Landscape, The Great Masturbator in Surrealist Landscape with DNA* (1957–8; Fig. 10) locates a prettified evocation of a space-filling model in one of Dali's typically barren landscapes inhabited by sub-Freudian enigmas, designed to conjure up a dreamworld of obscure sexual fantasy⁴. Subsequent artists, particularly those who have engaged with the social implication of molecular biology and genetic engineering, have located

images of DNA in contexts of meaning that are less obscure and more polemic.

This savagely selective glance at DNA art — omitting such contemporary luminaries of genetic art as Suzanne Anker⁵ (Fig. 11; www.geneculture.org), David Kremers



Figure 9 Paintings of DNA models on a 'Millennium Collection' stamp, designed by Mark Curtis (1999–2000), from the UK Royal Mail's Scientists' Tale collection.

MIRIAM CHUA, COLD SPRING HARBOR LAB.

ROYAL MAIL

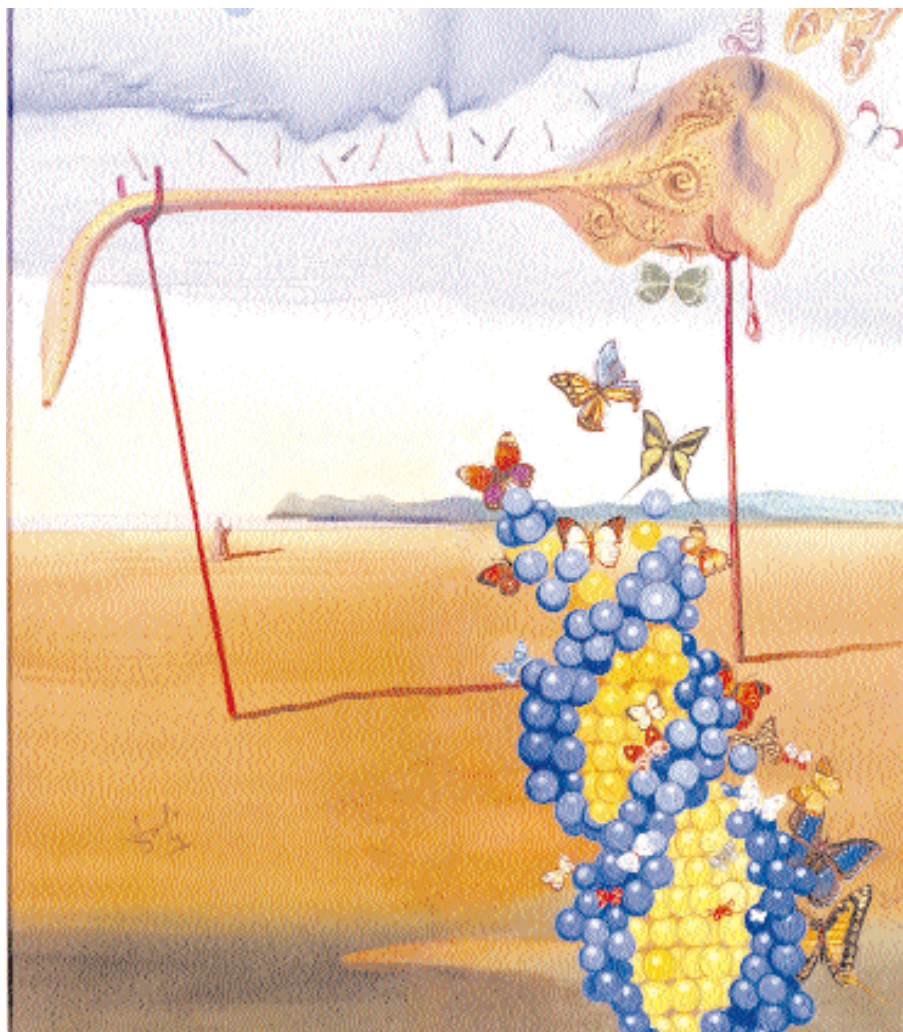


Figure 10 *Butterfly Landscape, The Great Masturbator in Surrealist Landscape with DNA* by Salvador Dalí, 1957–8. Private collection.

(<http://davidkremers.caltech.edu/>), Ellen Levy (<http://www.geneart.org/genome-levy.htm>), Sonya Rapoport (<http://users.lmi.net/sonyarap/transgenicbagel/>) and Gary Schneider (<http://www.icp.org/exhibitions/schneider/>) — can barely claim to be representative even of the main range of possibilities. In particular, exploitation of

the replicating potential of DNA to generate self-organizing images — exemplified by Marc Quinn's genetic portrait of Sir John Sulston from Sulston's own DNA, fragmented and replicated in bacterial colonies on plates of agar jelly⁶ — shows that some artists' engagement with DNA is maturing beyond iconographical opportunism.

Box 1

2003 exhibitions celebrating art in the age of the double helix

2003 is to throw up a series of exhibitions, including:

- Representations of the Double Helix, at the Whipple Museum of the History of Science, Department of History and Philosophy of Science, University of Cambridge, UK. The exhibition is curated by Soraya de Chadarevian and Harmke Kamminga, with the assistance of Corrina Bower, and will run from January to December 2003.
- Genetic Expressions: Art After DNA, at the Heckscher Museum of Art, Huntington, New York. Curated by Elizabeth Meryman and Lynn Gamwell, the exhibition will run from 28 June to 7 September 2003.
- From Code to Commodity: Genetics and Visual Art, at the New York Academy of Sciences. The exhibition runs from 13 February to 11 April 2003 and is curated by Dorothy Nelkin and Suzanne Anker. Cold Spring Harbor Laboratory Press are publishing a book by Nelkin and Anker entitled *The Molecular Gaze: Art in the Genetic Age*.
- PhotoGENESIS: Opus 2 — Artists' Response To the Genetic Information Age, at the Santa Barbara Museum of Art, 9 November 2002–9 February 2003.



Figure 11 *Zoosemiotics: Primates, Frog, Gazelle, Fish* (detail) by Suzanne Anker (1993).

But as with the *Mona Lisa*, opportunism will always be the name of a prominent public game. Typical of this tendency is the introduction by the perfume company Bijan in 1993 of a fragrance named DNA. Ironically, we learn from the maker's blurb that "DNA is recommended for casual use". Such is the destiny of one of the greatest popular icons. □

doi:10.1038/nature01403

1. Coen, E. *The Art of Genes. How Organisms Grow Themselves* (Oxford Univ. Press, Oxford, 1999).
2. Watson, J. D. & Crick, F. H. C. A structure for deoxyribonucleic acid. *Nature* **171**, 737–738 (1953).
3. Watson, J. D. *Molecular Biology of the Gene* (W. A. Benjamin, Inc., New York, 1965).
4. Descharnes, H. & Neret, G. *Salvador Dalí 1904–1989* (Taschen, New York, 1998).
5. Nelkin, D. & Anker, S. The influence of genetics on contemporary art. *Nature Rev. Genet.* **3**, 967–971 (2002).
6. Kemp, M. Reliquary and replication. *A Genomic Portrait: Sir John Sulston* by Marc Quinn. *Nature* **413**, 778 (2001).

Acknowledgements

R. Hodgson provided research assistance. Important advice has been gratefully received from S. de Chadarevian and H. Kamminga of the Department of the History and Philosophy of Science at Cambridge University, E. Levy, E. Coen, and K. Roberts of the John Innes Centre at Norwich.

Portrait of a molecule

Philip Ball

Nature, The Macmillan Building, 4 Crinan Street, London N1 9XW, UK (e-mail: p.ball@nature.com)

The double helix is idealized for its aesthetic elegant structure, but the reality of DNA's physical existence is quite different. Most DNA in the cell is compressed into a tangled package that somehow still exposes itself to meticulous gene-regulatory control. Philip Ball holds a mirror up to what we truly know about the mysteries of DNA's life inside a cell.

"Each level of organization represents a threshold where objects, methods and conditions of observation suddenly change. . . . Biology has then to articulate these levels two by two, to cross each threshold and unveil its peculiarities of integration and logic." François Jacob, in *The Logic of Life* (Penguin, London, 1989).

Rather like those of Albert Einstein, DNA's popular images are hardly representative. While it is fashionable in these post-genome days to show it as an endless string of A's, C's, G's and T's, this year's anniversary will surely be replete with two kinds of picture. One shows the famous double helix, delightfully suggesting the twin snakes of Wisdom and Knowledge intertwining around the caduceus, the staff of the medic's god Hermes. The other reveals the X-shaped symbol of inheritance, the chromosome.

But it is rare that DNA looks this good. For only a couple of hours during the early stages of the cell cycle, as the cell prepares to divide, the genome is compacted into its distinctive chromosomal fragments (Fig. 1). The rest of the time you will search the eukaryotic cell in vain for those molecular tetrapods. What you find instead in the cell nucleus is, apparently, a tangled mess.

And don't think that this will, on closer inspection, turn out to be woven from that elegant, pristine double helix. Rather, the threads are chromatin — a filamentary assembly of DNA and proteins — in which only very short stretches of the naked helix are fleetingly revealed. Although chromosomes are often equated with DNA, there is actually about twice as much protein as DNA in chromatin. And about 10 per cent of the mass of a chromosome is made up of RNA chains, newly formed (or in

the act of forming) on the DNA template in the process called transcription.

Zooming in on DNA

If we want to know how DNA really functions, it is not enough to zoom in to the molecular level with its beautifully simple staircase of base pairs. Textbooks tend, understandably, to show replication as the steady progress of the DNA-synthesizing enzyme DNA polymerase along a linear single strand laid out like a railway line (see article in this issue by Alberts, page 431), and RNA polymerase doing likewise in transcription. One has the impression of the genome as a book lying open, waiting to be read.

However, it is not so straightforward. The book is closed up, sealed, and packed away. Moreover, the full story is not merely what is written on the pages; these operations on DNA involve information transmission over

many length scales. Perhaps those who do not routinely have to delve into the intricacies of genome function have acquired such a simplistic picture of it all because, until relatively recently, these length scales were considered largely out of bounds for molecular science. We know about molecules; we know about cells and organelles; but the stuff in between is messy and mysterious.

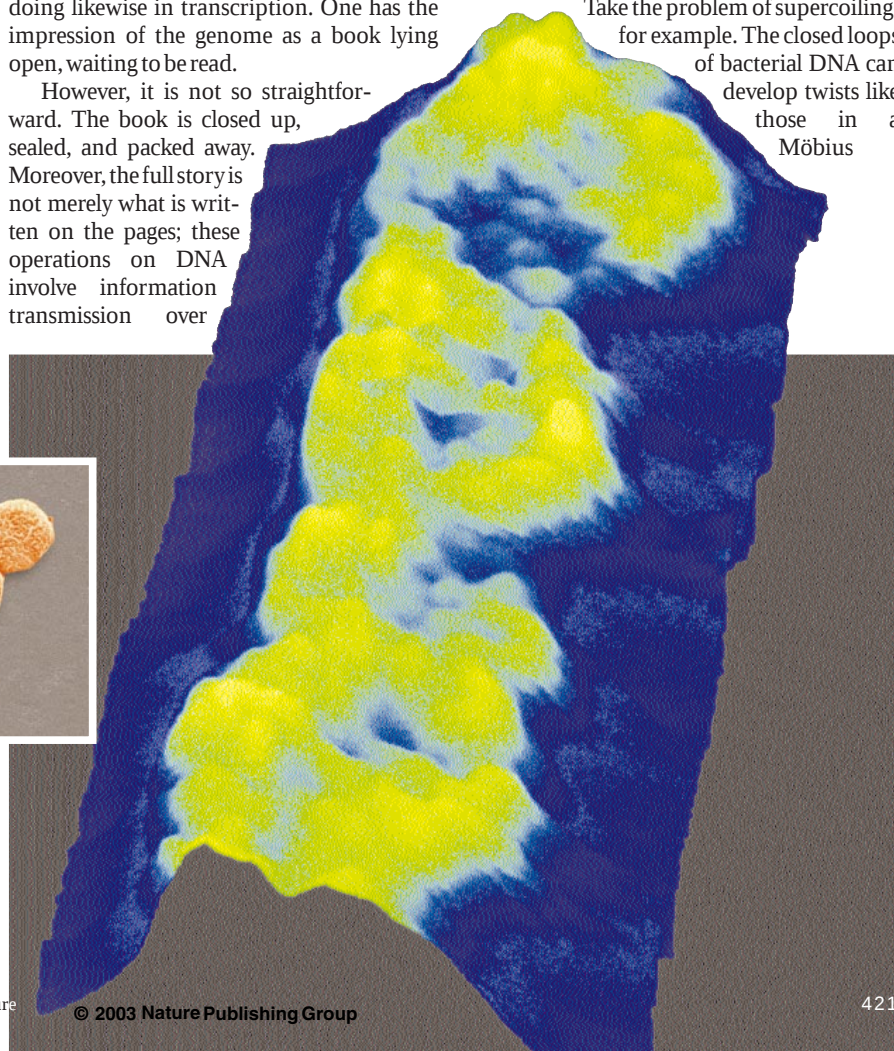
We speak of molecular biology and cell biology, but no one really talks of mesobiology. Yet that is the level of magnification at which much of the action takes place: the scale of perhaps a few to several hundred nanometres. How DNA is arranged on these scales seems to be central to the processes of replication and transcription that we have come to think of in terms of neat base pairings, yet it is precisely here that our understanding remains the most patchy.

Partly that's because the mesoscale represents, quite literally, a difficult middle ground. It encompasses too many atoms for one to apply straightforward molecular mechanics, with its bond bending and breaking; yet the graininess still matters, the continuum has not yet become a good approximation. As Bustamante and colleagues show elsewhere in this issue (see page 423), looking at DNA on a scale where it flexes and twists like a soft rod reveals how the mechanical and the molecular interact.

Take the problem of supercoiling, for example. The closed loops of bacterial DNA can develop twists like those in a Möbius



Figure 1 Inset: coloured scanning electron micrograph of a pair of human chromosomes. Main image: scanning tunnelling micrograph showing approximately three turns of a DNA double helix. The image is created by scanning a fine point just above the surface of a DNA molecule and electronically recording the height of the point as it moves across the specimen.



strip, which either 'overwind' or 'underwind' the helix. Generally there is some degree of underwinding (negative supercoiling) such that there is one negative supercoil for every 200 base pairs (bp) or so. This has an energy cost of around -9 kcal mol^{-1} , which manifests itself in physiological effects. In bacteria, too much supercoiling can inhibit growth, which is why enzymes called topoisomerases exist to release it. On the other hand, negative supercoiling tends to unwind the double helix, which is needed to initiate strand separation for DNA replication.

Although the chromosomal DNA of eukaryotes has free ends, it too is prone to supercoiling, as it seems typically to be attached in large loops to a filamentous structure called the nuclear matrix that coats the inside of the nuclear membrane. The attachment may in fact be necessary for both replication and transcription to take place.

Packaging problem

Stretched into a linear double helix, the three billion or so base pairs of human DNA would measure 1.8 m. This strand, snipped into 46 chromosomes, has to be packed into a nucleus just $6 \mu\text{m}$ or so across. As a result, the DNA chains are far from the idealized picture of molecules floating in an infinite solvent. They have a density of around 100 mg per ml, comparable to that of a highly viscous polymer gel.

The packing ratio for the chains is therefore enormous. In the smallest human chromosome, a length of DNA 14 mm long is compressed into a chromosome about $2 \mu\text{m}$ long: a packing ratio of 7,000. The first stage in solving this packaging problem is to wind the DNA around protein disks to form a bead-like nucleosome (see article in this issue by Felsenfeld and Groudine, page 448). Each disk is an octamer of four types of histone protein; a fifth histone, called H1, seals the DNA to the disk at the point where the winding starts and ends. Each nucleosome, 6 nm high by 11 nm in diameter, binds around 200 bp of DNA in two coils, and there is very little 'free' DNA between adjacent nucleosomes: sometimes as little as 8 bp.

The string of nucleosomes forms a fibre about 10 nm thick, which is then packaged into a filament three times as wide. This 30-nm fibre is the basic element of chromatin — yet we still don't know its structure. It is widely held to be composed of nucleosomes arranged in a solenoid, but hard evidence for this is scanty. How many celebrations of the double helix will admit that, 50 years on, we don't really know what DNA at large in the cell looks like?

The 30-nm fibre is further folded and condensed to give a packing ratio of around 1,000 in chromosomes during interphase (the time between cell divisions), and around ten times that in the X-shaped chromosomes of mitosis (cell division). How this happens is even more of a mystery. For mitotic chromosomes it

was thought until only recently that there might be a contiguous protein scaffold holding the whole affair together; but now it seems that the structural integrity must come from chromatin crosslinking¹. All the histones seem to have higher-order structural functions. Multi-subunit protein complexes in yeast called SWI/SNF and RSC (both of which seem to have human homologues) are chromatin-remodelling machines, which distort histone-DNA contacts or transfer histones between DNA molecules, exposing the DNA to attack by DNA-cleaving enzymes called nucleases. How they work remains hazy². According to one recent study³, DNA engaged by such complexes 'behaves as if it were free and bound at the same time'. Or in other words, as if 'free' and 'bound' were notions too simplistic to have much meaning here. What is clear is that these chromatin-shaping machines are important in transcription: cells lacking RSC are no longer viable.

There are in fact two types of chromatin in the nucleus of an interphase eukaryotic cell. Euchromatin is the most abundant: it is relatively dispersed, with the tangled-net appearance of a polymer gel. Heterochromatin is much denser (virtually solid-like), comparable to the density of mitotic chromosomes, and is confined to a few small patches. The invitation is to regard euchromatin as 'active' DNA, unpacked enough to let the transcription apparatus get to work on it, while heterochromatin is compressed, like a big data file, until needed. But like just about any other generalization about DNA's structure and behaviour, this one quickly breaks down. Clearly only a fraction of a cell's euchromatin is made up of transcribable DNA in the first place (so why not pack the rest away?); and even chromosomes containing a large amount of heterochromatin can be transcriptionally active. Some researchers think that 'euchromatin' and 'heterochromatin' are just blanket terms for many things we don't understand: further hierarchies of DNA organization yet to be revealed.

Structured chaos

Certainly, there seems to be more to the nucleus than a disorderly mass of DNA. It is a constantly changing structure, but not randomly: there is method in there somewhere. Specific chromosomes occupy discrete nuclear positions during interphase, and these positions can change in a deterministic way in response to changes in the cell's physiological state.

And the euchromatin itself has an internal logic, albeit one only partly decoded. It has been proposed that DNA has sequences called scaffold/matrix-attached regions (S/MARs), recurring typically every 10–100 kbp, that bind to the nuclear scaffold to divide up the chromosome into loops². Yet the existence of not only S/MARs, but also the nuclear scaffold itself, has been questioned. There is no sign of the scaffold during mitosis, and the material it is thought to be

composed of may be nothing more than a mess of denatured proteins.

Be that as it may, the organization of the loops seems to be important for compaction of DNA and for the regulation of gene expression, and each loop may act as an independent unit of gene activity. In other words, there is at least one level of superstructural organization in the chromosomes that makes its influence felt at the scale of molecular information transfer. Topoisomerase II is one of several proteins that bind specifically to the putative S/MARs, suggesting that these points are important for controlling supercoiling in the strands.

With all this high-level structure, transcription of DNA is not so much a matter of slotting the parts in place as tugging on the rope. DNA is highly curved around the nucleosomes, the inward-facing groove compressed and the outer one widened. RNA polymerase, at 13 by 14 nm, is about the same size as the nucleosome, yet it binds to a region of DNA around 50 bp long: about a quarter of the entire histone-bound length. So clearly some DNA must leave the surface of the histone core for transcription to proceed. But this core need not be displaced completely. The histone disk actually has a considerable amount of mobility, sometimes described as a corkscrew motion through the DNA coil. The reality is undoubtedly more complex, involving a kind of diffusion of localized defects in the DNA-histone contact.

If all of this destroys the pretty illusion created by the iconic model of Watson and Crick, it surely also opens up a much richer panorama. The fundamental mechanism of information transfer in nucleic acids — complementary base pairing — is so elegant that it risks blinding us to the awesome sophistication of the total process. These molecules do not simply wander up to one another and start talking. They must first be designated for that task, and must then file applications at various higher levels before permission is granted, forming a complex regulatory network (see article in this issue by Hood and Galas, page 444). For those who would like to control these processes, and those who seek to mimic them in artificial systems, the message is that the biological mesoscale, far from being a regime where order and simplicity descend into unpredictable chaos, has its own structures, logic, rules and regulatory mechanisms. This is the next frontier at which we will unfold the continuing story of how DNA works. □

doi:10.1038/nature01404

1. Poirier, M. G. & Marko, J. F. Mitotic chromosomes are chromatin networks without a mechanically contiguous protein scaffold. *Proc. Natl Acad. Sci. USA* **99**, 15393–15397 (2002).
2. Van Driel, R. & Otte, A. P. (eds) *Nuclear Organization, Chromatin Structure, and Gene Expression* (Oxford Univ. Press, Oxford, 1997).
3. Asturias, F. J., Chung, W.-H., Kornberg, R. D. & Lorch, Y. Structural analysis of the RSC chromatin-remodeling complex. *Proc. Natl Acad. Sci. USA* **99**, 13477–13480 (2002).

Ten years of tension: single-molecule DNA mechanics

Carlos Bustamante^{*†}, Zev Bryant^{*} & Steven B. Smith[†]

^{*}Department of Molecular and Cell Biology, and [†]Department of Physics and Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA (e-mail: carlos@alice.berkeley.edu; zev@alice.berkeley.edu; steve@alice.berkeley.edu)

The basic features of DNA were elucidated during the half-century following the discovery of the double helix. But it is only during the past decade that researchers have been able to manipulate single molecules of DNA to make direct measurements of its mechanical properties. These studies have illuminated the nature of interactions between DNA and proteins, the constraints within which the cellular machinery operates, and the forces created by DNA-dependent motors.

The physical properties of the DNA double helix are unlike those of any other natural or synthetic polymer. The molecule's characteristic base stacking and braided architecture lend it unusual stiffness: it takes about 50 times more energy to bend a double-stranded DNA (dsDNA) molecule into a circle than to perform the same operation on single-stranded DNA (ssDNA). Moreover, the phosphates in DNA's backbone make it one of the most highly charged polymers known.

The protein machinery involved in copying, transcribing and packaging DNA has adapted to exploit these unique physical properties (see article by Alberts, pages 431). For example, RNA polymerases (which synthesize RNA from a DNA template) and helicases (which unwind the double helix to provide single-stranded templates for polymerases) have evolved as motors capable of moving along torsionally constrained DNA molecules. DNA-binding proteins can use the polymer's electrostatic potential to cling to DNA while they diffuse along the molecule in search of their target sequences. Topoisomerases break and rejoin the DNA to relieve torsional strain that accumulates ahead of the replication fork.

During the past ten years, direct manipulation of single molecules of DNA has expanded our understanding of the mechanical interactions between DNA and proteins, following a pattern in which basic investigations of DNA elasticity have laid the groundwork for real-time, single-molecule assays of enzyme mechanism.

DNA as a worm-like chain

Although mechanical properties vary according to local sequence and helical structure, the relevant physics of DNA in many biological contexts is usefully described using a coarse-grained treatment such as the worm-like chain (WLC) model¹, which characterizes a polymer using a single parameter, the flexural persistence length (A). The WLC model imagines a polymer as a line that bends smoothly under the influence of random thermal fluctuations. The value of A defines the distance over which the direction of this line persists: correlation between the orientations of two polymer segments falls off exponentially (with decay length A) according to the contour length that separates them. For dsDNA in physiological buffer, $A \sim 50$ nm.

There is a simple relationship between A and the bending rigidity κ of the polymer represented as an elastic rod²: $k_B TA = \kappa$, where k_B is

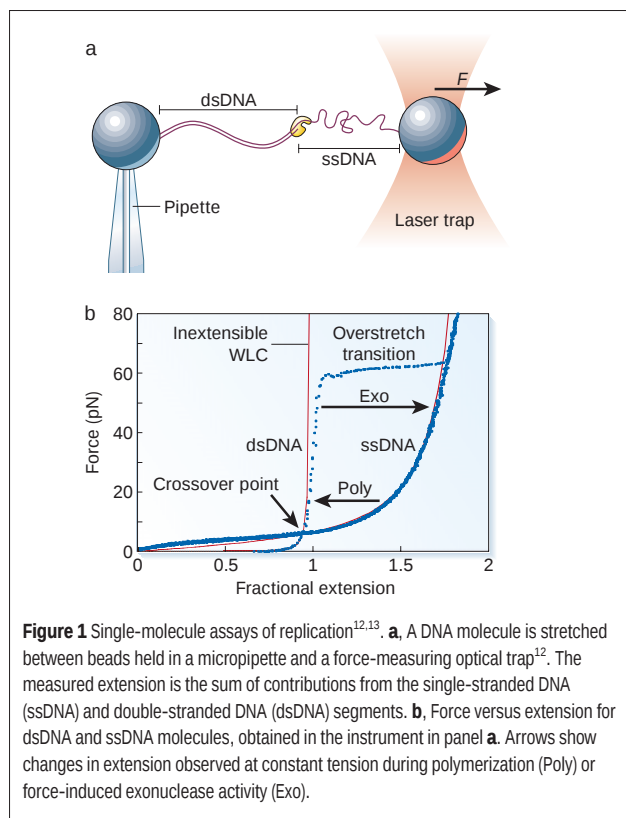


Figure 1 Single-molecule assays of replication^{12,13}. **a**, A DNA molecule is stretched between beads held in a micropipette and a force-measuring optical trap¹². The measured extension is the sum of contributions from the single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA) segments. **b**, Force versus extension for dsDNA and ssDNA molecules, obtained in the instrument in panel **a**. Arrows show changes in extension observed at constant tension during polymerization (Poly) or force-induced exonuclease activity (Exo).

Boltzmann's constant and T is the temperature. According to this relation, the energy required to bend a segment of DNA of length L through an angle θ and a radius of curvature R/L is:

$$E(\theta) = \frac{k_B T A L}{2R^2} = \frac{k_B T A}{2L} \theta^2$$

This model, therefore, predicts that it is energetically more favourable to bend the molecule smoothly, spreading the strain over large distances, than to bend it sharply at discrete locations. This mechanical property is central to interactions with regulatory proteins that bend DNA severely upon binding. The biological relevance of these bends is demonstrated by the enhancement of DNA recombination and gene transcription observed when specific protein-binding sites for activators are replaced by intrinsically bent DNA sequences³ or by binding sequences for unrelated DNA-bending proteins in the presence of these proteins⁴.

To bend DNA, proteins must convert part of their binding energy into mechanical work, as illustrated by an experiment in which a binding sequence was pre-bent towards the major groove by placing it in a DNA minicircle. The affinity of a transcription factor (TBP) for this binding site was found to be 300-fold higher (equivalent to a free-energy change of 3.4 kcal mol⁻¹) when the sequence was pre-bent in the same direction as TBP-induced bending, relative to pre-bending in the opposite direction⁵. This increase can largely be accounted for by the difference in bending energy between the two initial DNA conformations, which by the equation above is predicted to be 3.2 kcal mol⁻¹.

The high linear charge density of the double helix provides one mechanism for converting binding energy into work. DNA's structure is pre-stressed by electrostatic self-repulsion, as a result of the negatively charged phosphate backbone of the double helix. Therefore, asymmetric neutralization of the DNA helix (for example, by a DNA-binding protein that presents a positively charged face) can lead to compression and bending of DNA towards the neutralized face. This

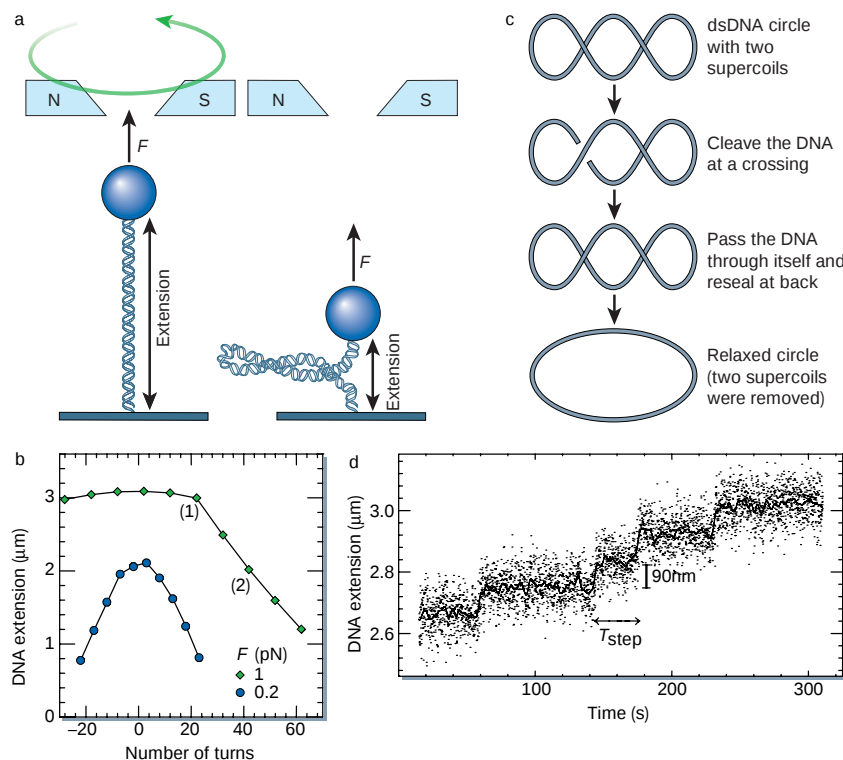


Figure 2 The elastic behaviour of supercoiled DNA molecules⁴⁹ forms the basis of single-molecule topoisomerase assays^{15–17}.

a, Molecules are stretched and twisted in magnetic tweezers. Under sufficient torsional strain, a twisted DNA molecule buckles to form plectonemes, shortening the measured extension. **b**, Extension as a function of turns introduced into the molecule remains nearly constant until the buckling transition is reached (1), after which the molecule contracts linearly (2). Underwinding at 1 pN does not cause reduced extension because the strained molecule forms alternative underwound structures (for example, melted DNA) in preference to buckling. **c**, The accepted model of type II topoisomerase action: the enzyme binds to supercoiled DNA, cleaves both strands, and passes the double helix through itself, leading to the removal of two turns. (Redrawn from ref. 50, with permission.) **d**, When topoisomerase II is added to a plectonemized molecule in the magnetic tweezers, 90-nm steps in extension are seen, corresponding to the removal of two turns¹⁵, as predicted by the model.

effect has been elegantly demonstrated by incorporating neutral phosphate analogues or tethered cations onto one face of a DNA molecule⁶.

DNA elasticity

The bending elasticity of DNA has consequences beyond short-range interactions with proteins: the WLC model explicitly connects local bending mechanics with the statistics of global conformations. Thus, a polymer with smaller bending rigidity tends to adopt a more compact random-coil structure.

This preference is reflected in the phenomenon of entropic elasticity, which is responsible for the elastic properties of common polymeric materials such as rubber⁷. A flexible polymer coils randomly in solution, resulting in an average end-to-end distance much shorter than its contour length. Pulling the molecule into a more extended chain is entropically unfavourable, as there are fewer possible conformations at longer extensions, with only a single possible conformation (a perfectly straight line) for maximum extension. The resulting entropic force increases as a random coil is pulled from the ends; tensions on the order of $k_B T/A$ (~ 0.1 pN for dsDNA or 5 pN for ssDNA) are required to extend the molecule significantly.

Direct measurements of force and extension on single molecules of DNA provide the most rigorous test to date of theories of entropic elasticity. When magnets and fluid flow⁸ and later optical traps^{9,10} were used to stretch DNA molecules attached to micron-scale beads (Fig. 1), the entropic force–extension behaviour of dsDNA was found to agree closely with the WLC model¹¹. Tensions of ~ 6 pN, within the range of forces exerted by characterized molecular motors, stretch dsDNA to $\sim 95\%$ of its contour length.

The intrinsic flexibility of ssDNA causes it to maintain very compact conformations, so that its extension per base pair is shorter than that of dsDNA for forces smaller than ~ 6 pN. At higher forces, however, the situation is reversed. A single strand is not constrained to follow a helical path, so it becomes nearly twice as long as dsDNA as it is pulled straight (Fig. 1).

From elasticity to enzymology

A quantitative appreciation of the different elastic properties of

ssDNA and dsDNA has allowed researchers to observe replication of single DNA molecules^{12,13}. In these studies, a molecule of ssDNA was stretched between two surfaces, and a DNA polymerase was allowed to replicate the stretched molecule at a given constant tension. As ssDNA was converted into dsDNA by the polymerase, replication could be followed in real time by monitoring the extension (below 6 pN) or contraction (above 6 pN) of the molecule (Fig. 1).

These studies showed that the rate-limiting step of DNA replication, which involves closing a structural ‘fingers’ domain of the enzyme, is sensitive to DNA tension and is capable of generating forces as high as 35 pN. Small forces can accelerate the enzyme’s activity, probably by helping it to stretch the proximal collapsed template strand into the correct geometry for polymerization. A surprising result was the induction of a strong exonuclease activity (removal of nucleotides) at tensions above 40 pN (ref. 12). This effect provides a novel assay to investigate the proofreading mechanism of DNA polymerases.

Studies of the force–extension behaviour of single supercoiled DNA molecules further illustrate the progression from elasticity measurements to enzymology (Fig. 2). DNA tethers were stretched between a surface and a magnetic bead that could be rotated using magnets¹⁴. Because the molecule was attached at each end through multiple linkages on both strands, rotation of the bead led to the build up of torsional strain in the molecule. Under tension, such a molecule behaves roughly like a twisted rubber tube: as turns are added, the extension remains nearly constant until a critical amount of torque accumulates and the tube buckles, trading twist for writhe to form plectonemes (units of supercoiled DNA that project out of the molecular axis), thus reducing its apparent extension with each subsequent turn. As tension is increased, so does the energetic penalty for buckling; therefore, more turns must be added to reach the buckling transition.

The activity of topoisomerase II, an enzyme that relaxes supercoils in eukaryotic cells, has been analysed on single plectonemized DNA tethers¹⁵. Under conditions of limiting ATP, discrete steps in extension were observed that were attributable to single enzymatic turnovers. The size of these steps corresponded to the removal of two turns, confirming the

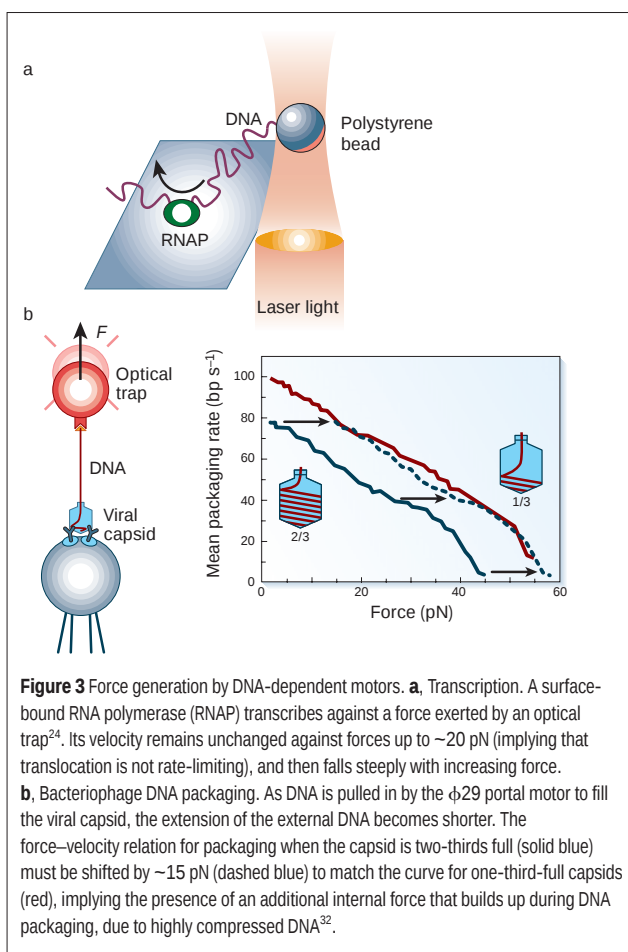


Figure 3 Force generation by DNA-dependent motors. **a**, Transcription. A surface-bound RNA polymerase (RNAP) transcribes against a force exerted by an optical trap²⁴. Its velocity remains unchanged against forces up to ~20 pN (implying that translocation is not rate-limiting), and then falls steeply with increasing force. **b**, Bacteriophage DNA packaging. As DNA is pulled in by the ϕ 29 portal motor to fill the viral capsid, the extension of the external DNA becomes shorter. The force–velocity relation for packaging when the capsid is two-thirds full (solid blue) must be shifted by ~15 pN (dashed blue) to match the curve for one-third-full capsids (red), implying the presence of an additional internal force that builds up during DNA packaging, due to highly compressed DNA³².

accepted model of topoisomerase II action in which the double helix is passed through itself, changing the linking number by two (Fig. 2). Later experiments applied the same methodology to bacterial topoisomerases I and IV (refs 16 and 17, respectively).

Single-molecule assays of topoisomerase IV (ref. 17) helped reveal that the enzyme has a chiral substrate specificity: it relaxes overwound DNA substantially more efficiently than underwound DNA. This allows the enzyme to relax the positive supercoils formed during replication while avoiding counterproductive relaxation of the negative supercoils present in non-replicating DNA.

These studies also showed that topoisomerase IV relaxes DNA (of either handedness) an order of magnitude faster than had been estimated from bulk studies, helping to resolve a dilemma in which the enzyme's apparent low turnover rate *in vitro* had seemed to be at odds with its demonstrated ability to counteract rapid supercoil formation at the replication fork *in vivo*. This result reflects a general caveat for bulk enzyme kinetics: the presence of inactive enzymes in solution can lead to gross underestimation of the turnover rate per enzyme. Single-molecule assays, including those based on DNA manipulation, can sidestep this issue by selecting only the active fraction for analysis.

DNA unzipped

DNA helicases must generate force to unzip the parental strands during replication (see article by Alberts, page 431). Mechanical unzipping forces for dsDNA were first measured by attaching one strand to a surface (through a dsDNA linker) and pulling on the other strand using a glass needle¹⁸ or optical tweezers¹⁹ as a force transducer. DNA from a bacterial virus, called lambda phage, was unzipped (and re-zipped) at forces between 10 and 15 pN, depend-

ing on the local sequence. The pattern of force fluctuations during unzipping of a particular sequence was remarkably reproducible, and could be rationalized from a simple model incorporating the known difference in stability between adenine–thymine and guanine–cytosine base pairs.

Future studies might use this experimental geometry to investigate helicase function directly. Such an experiment could provide insight into the rates, processivity, force generation and sequence dependence of helicases, complementing the results of previous single-molecule helicase assays which observed translocation without measuring or applying forces^{20,21}.

In a new application of mechanical unzipping²², the extra force needed to separate the DNA strands past DNA-binding proteins has been used to map the positions of target sequences, and (by noting the fraction of occupied sites as a function of protein concentration) to measure dissociation constants.

Forces in DNA transcription and packaging

The ability to apply forces on DNA has altered the way we think about DNA-dependent enzymes, by revealing these enzymes to be powerful motors (Fig. 3). Optical tweezers have been used to follow transcription by *Escherichia coli* RNA polymerase against external loads^{23–25}. This enzyme can generate forces exceeding those of cytoskeletal motors that drive transport processes within the cell²⁶. Its velocity remains unchanged against forces of up to ~20 pN (refs 24,25), showing that the translocation step (which must by definition be force sensitive) is not rate limiting.

An external load can, however, affect the tendency of an enzyme²⁵ to pause or arrest during transcription. The application of force in an 'aiding' direction²⁷ reduces pausing and arrest probabilities, presumably by preventing the polymerase from sliding backwards along the template during entry into these inactive states²⁸. The same 'backsliding' phenomenon may be responsible for the steep drop in transcriptional velocity seen as opposing force is increased above ~20 pN (ref. 24).

In eukaryotes, forces generated by RNA polymerase or by chromatin-remodelling enzymes might help to displace nucleosomes that would otherwise impede transcription. In support of this idea, several groups have pulled on single chromatin fibres and found that nucleosomes can be removed from DNA by applying a tension of ~20 pN (refs 29–31). At lower forces (~6 pN), reversible modifications of the chromatin structure (such as partial DNA release³¹ or disruption of internucleosomal interactions²⁹) are observed. These tension-inducible structural rearrangements might be exploited by RNA polymerase or other cellular factors to modulate access of the transcriptional apparatus to chromosomal DNA.

The machine that packs DNA into the viral capsid of the bacteriophage ϕ 29 (a virus that infects bacteria) generates higher forces than have been seen so far for any other translocating (displacing) molecular motor³² (Fig. 3). A comparison of the force–velocity relation of the motor when the capsid is mostly empty with that when it is nearly full of DNA revealed the presence of a large internal force (up to ~50 pN) pushing back on the motor. This pressure, which must be overcome in order to package the viral genome, presumably arises from the combined effects of configurational entropy loss, elastic bending energy, electrostatic self-repulsion, and changes in hydration of the DNA upon packaging. The potential energy thus stored by the motor in the form of a pre-compressed 'spring' should provide some of the driving force for DNA ejection into the bacterial cytoplasm when the virus infects.

Extreme forms of DNA

The helical structure of DNA is highly adaptable and can assume various forms³³. Although the helix of dsDNA is typically right-handed and extended in aqueous solution (B form), it can become shorter and wider (A form) in dehydrating solution. Molecules with specific base sequences (alternating purines and pyrimidines) easily assume the left-handed Z form, which is longer than B form and has reverse twist. Recently, single-molecule manipulation experiments

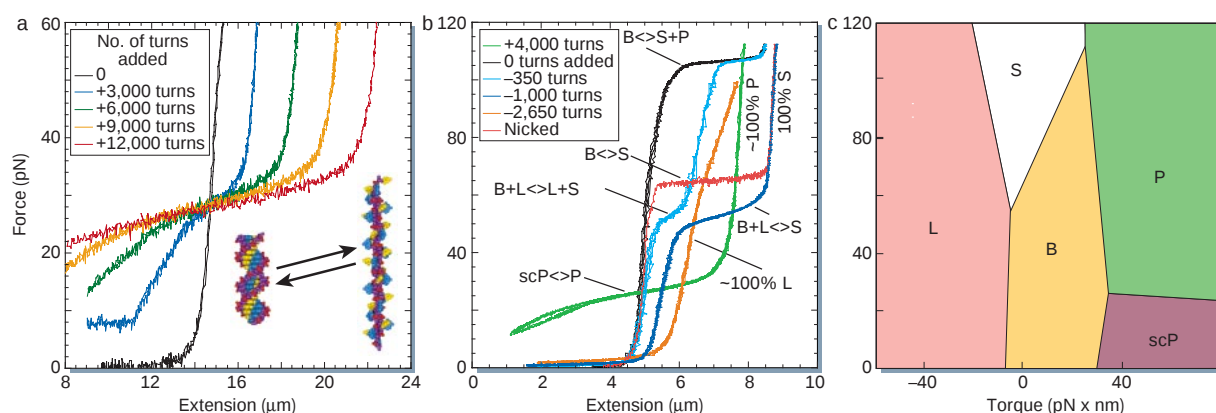


Figure 4 Mechanically induced structural transitions in twist-constrained DNA.

a, Force–extension curves for a 44.4-kilobase DNA molecule overwound by successively larger numbers of turns (Z.B., M. D. Stone, S.B.S., N. R. Cozzarelli and C.B., unpublished data). As was seen in ref. 39, the high-force curves can be interpreted as the sum of contributions from a fraction of the molecule that remains in B form and a progressively larger fraction that is converted to ‘P form’, whose structure (proposed from molecular mechanics modelling³⁹) is shown in the inset (courtesy of R. Lavery). The curves cross at an ‘isosbestic point’ marking the force at which B-DNA and P-DNA have equal extensions. **b**, Multiple plateaux occur in force–extension curves for torsionally constrained DNA³⁵. (Data shown are for a 14.8-kilobase molecule twisted

and stretched in 100 mM Na⁺; Z.B., M. D. Stone, S.B.S., N. R. Cozzarelli and C.B., unpublished data.) **c**, These multiple plateaux can be explained by a ‘phase diagram’ for DNA under torque and tension (adapted from ref. 41). Coloured regions represent conditions under which pure phases occur; lines indicate conditions for phase coexistence within a molecule. S, overstretched; P, Pauling, sc, supercoiled (shortened by forming plectonemes). L is used here in place of ‘Z’^{35,41} to denote a phase with an average left-handed twist. Other studies have concluded that this form contains exposed bases, consistent with melted DNA⁴⁸; a mixture of non-canonical forms may in fact be present. A nicked DNA molecule (red curve in **b**) remains at zero torque and therefore crosses the B–S coexistence line at 65 pN.

have revealed the existence of additional helical forms of DNA stabilized by external forces and torques (Fig. 4).

When tension in a nicked DNA molecule is increased to 65 pN, it displays a reversible, cooperative transition to an extended form that is ~70 per cent longer than normal B-DNA^{9,34} and with substantially reduced twist³⁵ (Figs 1b, 4b). But, what is the form of this overstretched dsDNA? Do the strands associate in some specific base-paired structure, dubbed ‘S form’^{9,34}, or does overstretched DNA simply comprise two independent strands of ssDNA³⁶? Evidence exists for both models, so the question remains open; a further challenge in single-molecule mechanics is the development of methods to probe the high-resolution structure of manipulated molecules³⁷.

Twisting of stretched DNA can lead to other structural transitions^{35,38,39}. For example, after a critical amount of overwinding has been introduced into a molecule (Fig. 4a), it gets progressively longer with additional twisting, implying cooperative conversion to an overextended form with greatly increased helicity (~2.6 base pairs per turn, compared with 10.5 base pairs per turn for B-DNA). The evidence³⁹ suggests an inside-out double helix reminiscent of the structure proposed by Linus Pauling in 1953 (ref. 40) and therefore dubbed Pauling DNA (P-DNA).

Complex force–extension curves with multiple force plateaux are seen when single DNA molecules are twisted in either direction and pulled to high forces (ref. 35 and Fig. 4b). A simple model to account for these features assumes that DNA has five interconvertible structural forms⁴¹. This model predicts a force–torque ‘phase diagram’ (Fig. 4c), thus framing mechanically induced structural transitions in terms of coexistence lines, critical stresses, and triple-points. Such a model might be tested by direct measurements of torque on stretched and twisted DNA — so far, this quantity has been inferred only indirectly from force–extension experiments. It remains to be determined whether molecular motors can generate sufficient concomitant torque and tension to generate ‘extreme forms’ of DNA in a biological context.

From mechanics to nanotechnology

Single-molecule manipulation of DNA has illuminated the mechanical basis of interactions between DNA and the molecular machinery

involved in transcription, replication and recombination. Over the next decade, these studies are likely to expand to include detailed analyses of the mechanical interactions of many factors involved in these fundamental cellular processes. Because of the potentially large class of motors that track the DNA helix (as demonstrated elegantly for RNA polymerase⁴²), necessary technical improvements will include direct measurement of torque in experiments that decouple twisting from bending.

Outside of traditional DNA biology, the ease of synthesis and well-characterized elasticity of DNA make it an ideal material for stiff molecular ‘handles’ to manipulate other molecules. So far, such handles have been used to mechanically unfold molecules of RNA⁴³, but covalent attachment of DNA segments to protein molecules has also been demonstrated⁴⁴, opening the door for the next generation of forced protein (un)folding studies⁴⁵ and perhaps mechanical assays of domain motion in enzymes.

Engineers have recently exploited the properties of DNA to construct self-assembled nanomachines, such as artificial DNA-based devices driven by strand displacement^{46,47} or chemically induced structural rearrangements (ref. 48; and see article by Seeman, page 427). DNA micromanipulation techniques will help assess the utility of this new class of molecular machines for which force and torque generation have yet to be measured. The past decade has provided a new perspective of the mechanical nature of the double helix. The next decade promises deeper insight into its interactions with the cellular machinery and its potential for constructing sophisticated nanomachines. □

doi:10.1038/nature01405

- Kratky, O. & Porod, G. Röntgenuntersuchung gelöster Fadenmoleküle. *Rec. Trav. Chim. Pays-Bas* **68**, 1106–1123 (1949).
- Schellman, J. A. Flexibility of DNA. *Biopolymers* **13**, 217–226 (1974).
- Goodman, S. D. & Nash, H. D. Functional replacement of a protein-induced bend in a DNA recombination site. *Nature* **341**, 251–254 (1989).
- Perez-Martin, J. & Espinosa, M. Protein-induced bending as a transcriptional switch. *Science* **260**, 805–807 (1993).
- Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature* **373**, 724–727 (1995).
- Strauss, J. K. & Maher, L. J. III DNA bending by asymmetric phosphate neutralization. *Science* **266**, 1829–1834 (1994).

DNA in a material world

Nadrian C. Seeman

Department of Chemistry, New York University, New York 10003, USA
(e-mail: ned.seeman@nyu.edu)

The specific bonding of DNA base pairs provides the chemical foundation for genetics. This powerful molecular recognition system can be used in nanotechnology to direct the assembly of highly structured materials with specific nanoscale features, as well as in DNA computation to process complex information. The exploitation of DNA for material purposes presents a new chapter in the history of the molecule.

"The nucleic-acid 'system' that operates in terrestrial life is optimized (through evolution) chemistry incarnate. Why not use it ... to allow human beings to sculpt something new, perhaps beautiful, perhaps useful, certainly unnatural." Roald Hoffmann, writing in *American Scientist*, 1994 (ref. 1).

The DNA molecule has appealing features for use in nanotechnology: its minuscule size, with a diameter of about 2 nanometres, its short structural repeat (helical pitch) of about 3.4–3.6 nm, and its 'stiffness', with a persistence length (a measure of stiffness) of around 50 nm. There are two basic types of nanotechnological construction: 'top-down' systems are where microscopic manipulations of small numbers of atoms or molecules fashion elegant patterns (for example, see ref. 2), while in 'bottom-up' constructions, many molecules self-assemble in parallel steps, as a function of their molecular recognition properties. As a chemically based assembly system, DNA will be a key player in bottom-up nanotechnology.

The origins of this approach date to the early 1970s, when *in vitro* genetic manipulation was first performed by tacking together molecules with 'sticky ends'. A sticky end is a short single-stranded overhang protruding from the end of a double-stranded helical DNA molecule. Like flaps of Velcro, two molecules with complementary sticky ends — that is, their sticky ends have complementary arrangements of the nucleotide bases adenine, cytosine, guanine and thymine — will cohere to form a molecular complex.

Sticky-ended cohesion is arguably the best example of programmable molecular recognition: there is significant diversity to possible sticky ends (4^N for N -base sticky ends), and the product formed at the site of this cohesion is the classic DNA double helix. Likewise, the convenience of solid support-based DNA synthesis³ makes it easy to program diverse sequences of sticky ends. Thus, sticky ends offer both predictable control of intermolecular associations and predictable geometry at the point of cohesion. Perhaps one could get similar affinity properties from antibodies and antigens, but, in contrast to DNA sticky ends, the relative three-dimensional orientation of the antibody and the antigen would need to be determined for every new pair. The nucleic acids seem to be unique in this regard, providing a tractable, diverse and programmable system with remarkable control over intermolecular interactions, coupled with known structures for their complexes.

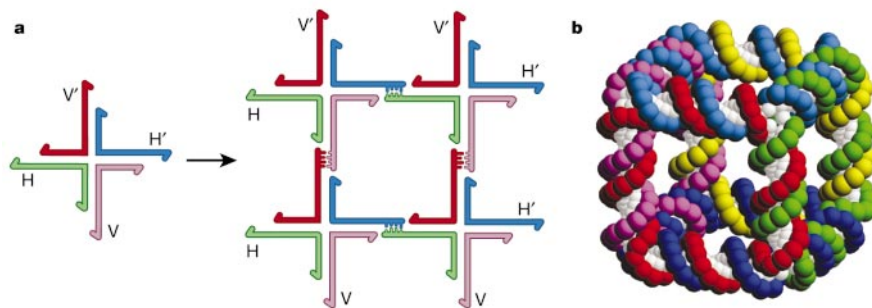
Branched DNA

There is, however, a catch; the axes of DNA double helices are unbranched lines. Joining DNA molecules by sticky ends can yield longer lines, perhaps with specific components in a particular linear

- Beuche, F. *Physical Properties of Polymers* (Interscience, New York, 1962).
- Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
- Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271**, 795–799 (1996).
- Wang, M. D., Yin, H., Landick, R., Gelles, J. & Block, S. M. Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997).
- Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599–1600 (1994).
- Wuite, G. J., Smith, S. B., Young, M., Keller, D. & Bustamante, C. Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature* **404**, 103–106 (2000).
- Maier, B., Bensimon, D. & Croquette, V. Replication by a single DNA polymerase of a stretched single-stranded DNA. *Proc. Natl Acad. Sci. USA* **97**, 12002–12007 (2000).
- Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. & Croquette, V. The elasticity of a single supercoiled DNA molecule. *Science* **271**, 1835–1837 (1996).
- Strick, T. R., Croquette, V. & Bensimon, D. Single-molecule analysis of DNA uncoiling by a type II topoisomerase. *Nature* **404**, 901–904 (2000).
- Dekker, N. H. *et al.* The mechanism of type IA topoisomerases. *Proc. Natl Acad. Sci. USA* **99**, 12126–12131 (2002).
- Crisona, N. J., Strick, T. R., Bensimon, D., Croquette, V. & Cozzarelli, N. R. Preferential relaxation of positively supercoiled DNA by *E. coli* topoisomerase IV in single-molecule and ensemble measurements. *Genes Dev.* **14**, 2881–2892 (2000).
- Essevez-Roulet, B., Bockelmann, U. & Heslot, F. Mechanical separation of the complementary strands of DNA. *Proc. Natl Acad. Sci. USA* **94**, 11935–11940 (1997).
- Bockelmann, U., Thomen, P., Essevez-Roulet, B., Viasnoff, V. & Heslot, F. Unzipping DNA with optical tweezers: high sequence sensitivity and force flips. *Biophys. J.* **82**, 1537–1553 (2002).
- Bianco, P. R. *et al.* Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. *Nature* **409**, 374–378 (2001).
- Dohoney, K. M. & Gelles, J. χ -Sequence recognition and DNA translocation by single RecBCD helicase/nuclease molecules. *Nature* **409**, 370–374 (2001).
- Koch, S. J., Shundrovsky, A., Jantzen, B. C. & Wang, M. D. Probing protein-DNA interactions by unzipping a single DNA double helix. *Biophys. J.* **83**, 1098–1105 (2002).
- Yin, H. *et al.* Transcription against an applied force. *Science* **270**, 1653–1657 (1995).
- Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282**, 902–907 (1998).
- Davenport, R. J., Wuite, G. J., Landick, R. & Bustamante, C. Single-molecule study of transcriptional pausing and arrest by *E. coli* RNA polymerase. *Science* **287**, 2497–2500 (2000).
- Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
- Forde, N. R., Izhaky, D., Woodcock, G. R., Wuite, G. J. & Bustamante, C. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *Escherichia coli* RNA polymerase. *Proc. Natl Acad. Sci. USA* **99**, 11682–11687 (2002).
- Landick, R. RNA polymerase slides home: pause and termination site recognition. *Cell* **88**, 741–744 (1997).
- Cui, Y. & Bustamante, C. Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl Acad. Sci. USA* **97**, 127–132 (2000).
- Bennink, M. L. *et al.* Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers. *Nature Struct. Biol.* **8**, 606–610 (2001).
- Brower-Toland, B. D. *et al.* Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA. *Proc. Natl Acad. Sci. USA* **99**, 1960–1965 (2002).
- Smith, D. E. *et al.* The bacteriophage ϕ 29 portal motor can package DNA against a large internal force. *Nature* **413**, 748–752 (2001).
- Calladine, C. R. & Drew, H. *Understanding DNA* (Academic, London, 1997).
- Cluzel, P. *et al.* DNA: an extensible molecule. *Science* **271**, 792–794 (1996).
- Leger, J. F. *et al.* Structural transitions of a twisted and stretched DNA molecule. *Phys. Rev. Lett.* **83**, 1066–1069 (1999).
- Williams, M. C., Rouzina, I. & Bloomfield, V. A. Thermodynamics of DNA interactions from single molecule stretching experiments. *Acc. Chem. Res.* **35**, 159–166 (2002).
- Wilkins, M. H. F., Gosling, R. G. & Seeds, W. E. Nucleic acid: an extensible molecule? *Nature* **167**, 759–760 (1951).
- Strick, T. R., Allemand, J. F., Bensimon, D. & Croquette, V. Behavior of supercoiled DNA. *Biophys. J.* **74**, 2016–2028 (1998).
- Allemand, J. F., Bensimon, D., Lavery, R. & Croquette, V. Stretched and overwound DNA forms a Pauling-like structure with exposed bases. *Proc. Natl Acad. Sci. USA* **95**, 14152–14157 (1998).
- Pauling, L. & Corey, R. B. A proposed structure for the nucleic acids. *Proc. Natl Acad. Sci. USA* **39**, 84–97 (1953).
- Sarkar, A., Leger, J. F., Chatenay, D. & Marko, J. F. Structural transitions in DNA driven by external force and torque. *Phys. Rev. E* **63**, 051903–1–051903–10 (2001).
- Harada, Y. *et al.* Direct observation of DNA rotation during transcription by *Escherichia coli* RNA polymerase. *Nature* **409**, 113–115 (2001).
- Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. J. & Bustamante, C. Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**, 733–737 (2001).
- Hegner, M. DNA Handles for single molecule experiments. *Single Mol.* **1**, 139–144 (2000).
- Carrión-Vázquez, M. *et al.* Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog. Biophys. Mol. Biol.* **74**, 63–91 (2000).
- Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605–608 (2000).
- Yan, H., Zhang, X., Shen, Z. & Seeman, N. C. A robust DNA mechanical device controlled by hybridization topology. *Nature* **415**, 62–65 (2002).
- Mao, C., Sun, W., Shen, Z. & Seeman, N. C. A nanomechanical device based on the B–Z transition of DNA. *Nature* **397**, 144–146 (1999).
- Strick, T., Allemand, J., Croquette, V. & Bensimon, D. Twisting and stretching single DNA molecules. *Prog. Biophys. Mol. Biol.* **74**, 115–140 (2000).
- Alberts, B. *et al.* *Molecular Biology of the Cell* (Garland, New York, 2002).

Figure 1 Assembly of branched DNA molecules.

a, Self-assembly of branched DNA molecules into a two-dimensional crystal. A DNA branched junction forms from four DNA strands; those strands coloured green and blue have complementary sticky-end overhangs labelled H and H', respectively, whereas those coloured pink and red have complementary overhangs V and V', respectively. A number of DNA branched junctions cohere based on the orientation of their complementary sticky ends, forming a square-like unit with unpaired sticky ends on the outside, so more units could be added to produce a two-dimensional crystal. **b**, Ligated DNA molecules form interconnected rings to create a cube-like structure. The structure consists of six cyclic interlocked single strands, each linked twice to its four neighbours, because each edge contains two turns of the DNA



double helix. For example, the front red strand is linked to the green strand on the right, the light blue strand on the top, the magenta strand on the left, and the dark blue strand on the bottom. It is linked only indirectly to the yellow strand at the rear.

or cyclic order in one dimension. Indeed, the chromosomes packed inside cells exist as just such one-dimensional arrays. But to produce interesting materials from DNA, synthesis is required in multiple dimensions and, for this purpose, branched DNA is required.

Branched DNA occurs naturally in living systems, as ephemeral intermediates formed when chromosomes exchange information during meiosis, the type of cell division that generates the sex cells (eggs and sperm). Prior to cell division, homologous chromosomes pair, and the aligned strands of DNA break and literally cross over one another, forming structures called Holliday junctions. This exchange of adjacent sequences by homologous chromosomes — a process called recombination — during the formation of sex cells passes genetic diversity onto the next generation.

The Holliday junction contains four DNA strands (each member of a pair of aligned homologous chromosomes is composed of two DNA strands) bound together to form four double-helical arms flanking a branch point (Fig. 1a). The branch point can relocate throughout the molecule, by virtue of the homologous sequences. In contrast, synthetic DNA complexes can be designed to have fixed branch points containing between three and at least eight arms^{4,5}. Thus, the prescription for using DNA as the basis for complex materials with nanoscale features is simple: take synthetic branched DNA molecules with programmed sticky

ends, and get them to self-assemble into the desired structure, which may be a closed object or a crystalline array (Fig. 1a).

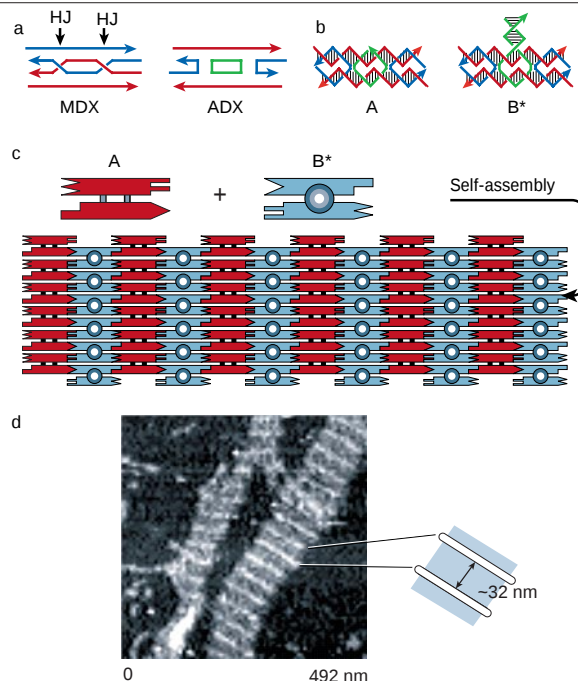
Other modes of nucleic acid interaction aside from sticky ends are available. For example, Tecto-RNA molecules⁶, held together by loop-loop interactions, or paranemic crossover (PX) DNA, where cohesion derives from pairing of alternate half turns in inter-wrapped double helices⁷. These new binding modes represent programmable cohesive interactions between cyclic single-stranded molecules that do not require cleavage to expose bases to pair molecules together. Nevertheless, cohesion using sticky ends remains the most prominent intermolecular interaction in structural DNA nanotechnology.

DNA constructions

It is over a decade since the construction of the first artificial DNA structure, a stick-cube, whose edges are double helices⁸ (Fig. 1b). More complex polyhedra and topological constructs⁹, such as knots and Borromean rings (consisting of three intricately interlinked circles), followed. But the apparent floppiness of individual branched junctions led to a hiatus before the next logical step: self-assembly into two-dimensional arrays.

This step required a stiffer motif, as it was difficult to build a periodic well-structured array with marshmallow-like components,

Figure 2 Two-dimensional DNA arrays. **a**, Schematic drawings of DNA double crossover (DX) units. In the meiotic DX recombination intermediate, labelled MDX, a pair of homologous chromosomes, each consisting of two DNA strands, align and cross over in order to swap equivalent portions of genetic information; 'HJ' indicates the Holliday junctions. The structure of an analogue unit (ADX), used as a tiling unit in the construction of DNA two-dimensional arrays, comprises two red strands, two blue crossover strands and a central green crossover strand. **b**, The strand structure and base pairing of the analogue ADX molecule, labelled A, and a variant, labelled B*. B* contains an extra DNA domain extending from the central green strand that, in practice, protrudes roughly perpendicular to the plane of the rest of the DX molecule. **c**, Schematic representations of A and B* where the perpendicular domain of B* is represented as a blue circle. The complementary ends of the ADX molecules are represented as geometrical shapes to illustrate how they fit together when they self-assemble. The dimensions of the resulting tiles are about 4 × 16 nm and are joined together so that the B* protrusions lie about 32 nm apart. **d**, The B* protrusions are visible as 'stripes' in tiled DNA arrays under an atomic force microscope.



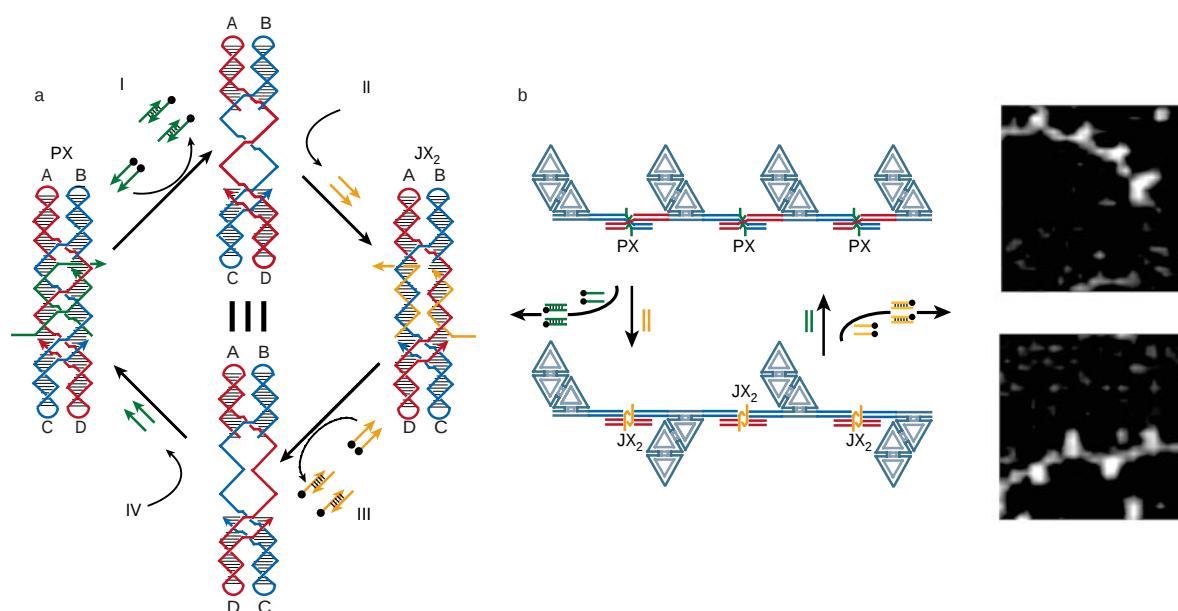


Figure 3 A rotary DNA nanomachine. **a**, The device works by producing two different conformations, depending on which of two pairs of strands (called 'set' strands) binds to the device framework. The device framework consists of two DNA strands (red and blue) whose top and bottom double helices are each connected by single strands. Thus, they form two rigid arms with a flexible hinge in between and the loose ends of the two strands dangling freely. The two states of the device, PX (left) and JX₂ (right), differ by a half turn in the relative orientations of their bottom helices (C and D on the left, D and C on the right). The difference between the two states is analogous to two adjacent fingers extended,

parallel to each other (right), or crossed (left). The states are set by the presence of green or yellow set strands, which bind to the frame in different ways to produce different conformations. The set strands have extensions that enable their removal when complementary strands are added (steps I and III). When one type of set strand is removed, the device is free to bind the other set strands and switch to a different state (steps II and IV). **b**, The PX–JX₂ device can be used to connect 20-nm DNA trapezoid constructs. In the PX state, they are in a parallel conformation, but in the JX₂ state, they are in a zig-zag conformation, which can be visualized on the right by atomic force microscopy.

even with a well-defined blueprint (sticky-ended specificity) for their assembly. The stiffer motif was provided by the DNA double-crossover (DX) molecule¹⁰, analogous, once again, to the double Holliday-junction intermediate formed during meiosis (MDX, Fig. 2a). This stiff molecule contains two double helices connected to each other twice through crossover points. It is possible to program DX molecules to produce a variety of patterned two-dimensional arrays just by controlling their sticky ends^{11–13} (Fig. 2b).

DNA nanomachines

In addition to objects and arrays, a number of DNA-based nanomechanical devices have been made. The first device consisted of two DX molecules connected by a shaft with a special sequence that could be converted from normal right-handed DNA (known as B-DNA) to an unusual left-handed conformation, known as Z-DNA¹⁴. The two DX molecules lie on one side of the shaft before conversion and on opposite sides after conversion, which leads to a rotation. The problem

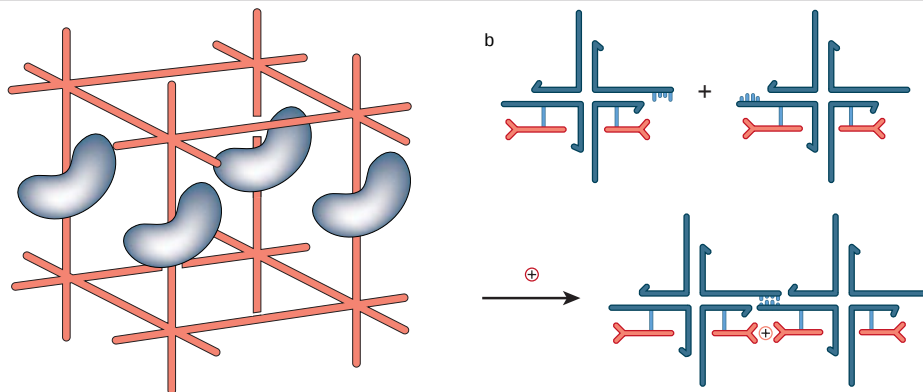
with this device is that it is activated by a small molecule, $\text{Co}(\text{NH}_3)_6^{3+}$, and with all devices sharing the same stimulus, an ordered collection of DX molecules would not produce a diversity of responses.

This problem was solved by Bernard Yurke and colleagues, who developed a protocol for a sequence-control device that has a tweezers-like motion¹⁵. The principle behind the device is that a so-called 'set' strand containing a non-pairing extension hybridizes to a DNA-paired structural framework and sets a conformation; another strand that is complementary to the 'set' strand is then added, which binds to both the pairing and non-pairing portions, and removes it from the structure, leaving only the framework.

A robust rotary device was developed based on this principle¹⁶ (Fig. 3), in which different set strands can enter and set the conformation to different structural end-states. In this way, the conformation of the DNA device can readily be flipped back and forth simply by adding different set strands followed by their complements. A variety of different devices can be controlled by a diverse group of set strands.

Figure 4 Applications of DNA scaffolds.

a, Scaffolding of biological macromolecules. A DNA box (red) is shown with protruding sticky ends that are used to organize boxes into crystals. Macromolecules are organized parallel to each other within the box, rendering them amenable to structure determination by X-ray crystallography. **b**, DNA scaffolds to direct the assembly of nanoscale electrical circuits. Branched DNA junctions (blue) direct the assembly of attached nanoelectronic components (red), which are stabilized by the addition of a positively charged ion.



DNA as a scaffold

What is the purpose of constructing DNA arrays and nanodevices? One prominent goal is to use DNA as scaffolding to organize other molecules. For example, it may be possible to use self-assembled DNA lattices (crystals) as platforms to position biological macromolecules so as to study their structure by X-ray crystallography⁴ (Fig. 4a). Towards this goal, programming of DNA has been used to bring protein molecules in proximity with each other to fuse multiple enzymatic activities¹⁷. However, the potential of this approach awaits the successful self-assembly of three-dimensional crystals.

Another goal is to use DNA crystals to assemble nanoelectronic components in two- or three-dimensional arrays¹⁸ (Fig. 4b). DNA has been shown to organize metallic nanoparticles as a precursor to nanoelectronic assembly^{19–22}, but so far it has not been possible to produce multidimensional arrays containing nanoelectronic components with the high-structural order of the naked DNA arrays described earlier.

There has been some controversy over whether DNA can be used as an electrical conductor (for example, ref. 23), although the resolution of this debate is unlikely have any impact on the use of DNA as a scaffold. Recently, the effects of DNA conformational changes on conduction in the presence of an analyte were shown to have potential as a biosensor²⁴.

Replicating DNA components

A natural question to ask of any assembly system based on DNA is whether the components can be replicated. To produce branched DNA molecules whose branch points do not move, they must have different sequences in opposite branches but, as a consequence, these structures are not readily reproduced by DNA polymerase; the polymerase would produce complements to all strands present, leading only to double helical molecules. One option is to use topological tricks to convert structures like the DNA cube into a long single strand by adding extra stretches of DNA bases. The single strand could then be replicated by DNA polymerase and the final replicated product induced to fold into the original shape, with any extraneous segments cleaved using restriction enzymes. Although this would produce a molecule with sticky ends ready to participate in self-assembly, it would be a cumbersome process²⁵.

Günter von Kiedrowski and colleagues have recently developed a way of replicating short, simple DNA branches in a mixed organic–DNA species. Their branched molecule consists of three DNA single strands bonded to an organic triangle-shaped linker. To replicate the branched molecule, the single-stranded complement of each of these strands is bound to the molecule, so that one end of each complement molecule is close to the same end of the other complement molecule. In the final step, the juxtaposed complements are connected together by bonding their neighbouring ends to another molecule of the organic linker²⁶. Extension of this system to the next level, such as objects like the cube, will need to solve topological problems involved in the separation of the two components, or it will be limited to unligated systems.

Future prospects

Many separate capabilities of DNA nanotechnology have been prototyped — it is now time to extend and integrate them into useful systems. Combining sequence-dependent devices with nanoscale arrays will provide a system with a vast number of distinct, programmable structural states, the *sine qua non* of nanorobotics. A key step in realizing these goals is to achieve highly ordered three-dimensional arrays, both periodic and, ultimately, algorithmic.

Interfacing with top-down nanotechnology will extend markedly the capabilities of the field. It also will be necessary to integrate biological macromolecules or other macromolecular complexes into DNA arrays in order to make practical systems with nanoscale components. Likewise, the inclusion of electronic components in highly ordered arrays will enable the organization of nanoelectronic circuits. Chemical function could be added to DNA arrays by adding nucleic acid species evolved *in vitro* to have specific binding properties ('aptamers') or enzymatic activities ('ribozymes' or 'DNAzymes'). A further area that has yet to have an impact on DNA nanotechnology is

Box 1

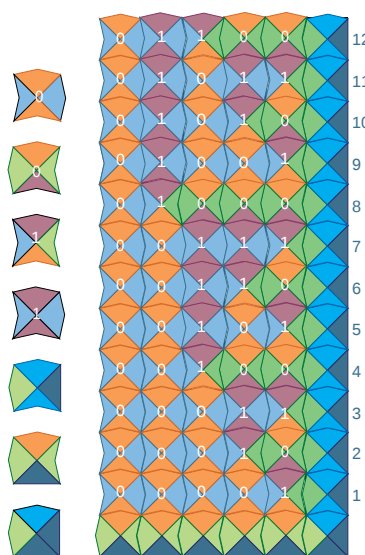
DNA computers

An assembly of DNA strands can process data in a similar way as an electronic computer, and has the potential to solve far more complex problems and store a greater amount of information, for substantially less energy costs than do electronic microprocessors. DNA-based computation dates from Leonard Adleman's landmark report in 1994 (ref. 27), where he used DNA to solve the 'Hamiltonian path' problem, a variant of the 'travelling salesman' problem. The idea is to establish whether there is a path between two cities, given an incomplete set of available roads. Adleman used strands of DNA to represent cities and roads, and encoded the sequences so that a strand representing a road would connect (according to the rules of base pairing) to any two strands representing a city. By mixing together the strands, joining the cities connected by roads, and weeding out any 'wrong answers', he showed that the strands could self-assemble to solve the problem.

It is impossible to separate DNA nanotechnology from DNA-based computation: many researchers work in both fields and the two communities have a symbiotic relationship. The first link between DNA computation and DNA nanotechnology was established by Erik Winfree, who suggested that short branched DNA molecules could be 'programmed' to undergo algorithmic self-assembly and thus serve as the basis of computation²⁸.

Periodic building blocks of matter, such as the DNA molecules shown in Fig. 1a, represent the simplest algorithm for assembly. All components are parallel, so what is on one side of a component is also on the other side, and in every direction. Given this parallelism, if the right side complements the left, the top complements the bottom and the front complements the back, a crystal should result. Even more complex algorithms are possible if one uses components of the same shape, but with different sticky ends. For example, Winfree has shown that, in principle, DNA tiles can be used to 'count' (see figure below) by creating borders with programmable sizes for one-, two- and possibly three-dimensional assemblies²⁹. If this scheme can be realized, self-assembly of precisely sized nanoscale arrays will be possible. A computation using self-assembly has been prototyped in one dimension, thereby lending some credence to the viability of algorithmic assembly³⁰.

Box 1 Figure Counting with self-assembled DNA tiles. DNA tiles are represented by squares with coloured edges that are protruded or indented. Seven component tiles are shown on the left: three border tiles on the bottom and four tiles with the values 0 or 1. The array illustrates binary counting from 1 (bottom row) to 12 (top row). Assembly is assumed to proceed by forming the reverse L-shaped border first, followed by binding the tiles that fit into the sites containing two (but not one) edges. Thus, the border determines the 1 tile in its bend, then that 1 tile and the horizontal-border tile on its left determine the 0 tile that fits, while the 1 tile and the vertical-border tile above it determine the (different) 0 tile that fits. (Adapted from ref. 29.)



combinatorial synthesis, which may well lead to greater diversity of integrated components. DNA-based computation and algorithmic assembly is another active area of research, and one that is impossible to separate from DNA nanotechnology (see Box 1).

The field of DNA nanotechnology has attracted an influx of researchers over the past few years. All of those involved in this area have benefited from the biotechnology enterprise that produces DNA-modifying enzymes and unusual components for synthetic DNA molecules. It is likely that applications in structural DNA nanotechnology ultimately will use variants on the theme of DNA (for example, peptide nucleic acids, containing an unconventional synthetic peptide backbone and nucleic acid bases for side chains), whose properties may be better suited to particular types of applications.

For the past half-century, DNA has been almost exclusively the province of biologists and biologically oriented physical scientists, who have studied its biological impact and molecular properties. During the next 50 years, it is likely they will be joined by materials scientists, nanotechnologists and computer engineers, who will exploit DNA's chemical properties in a non-biological context. □

doi:10.1038/nature01406

- Hoffmann, R. DNA as clay. *Am. Sci.* **82**, 308–311 (1994).
- Cuberes, M. T., Schlittler, R. R. & Gimzewski, J. K. Room-temperature repositioning of individual C-60 molecules at Cu steps: operation of a molecular counting device. *Appl. Phys. Lett.* **69**, 3016–3018 (1996).
- Caruthers, M. H. Gene synthesis machines: DNA chemistry and its uses. *Science* **230**, 281–285 (1985).
- Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **99**, 237–247 (1982).
- Seeman, N. C. Molecular craftwork with DNA. *Chem. Intell.* **1**, 38–47 (1995).
- Jaeger, L., Westhof, E. & Leontis, N. B. Tecto-RNA: modular assembly units for the construction of RNA nano-objects. *Nucleic Acids Res.* **29**, 455–463 (2001).
- Zhang, X., Yan, H., Shen, Z. & Seeman, N. C. Paranemic cohesion of topologically-closed DNA molecules. *J. Am. Chem. Soc.* **124**, 12940–12941 (2002).
- Chen, J. & Seeman, N. C. The synthesis from DNA of a molecule with the connectivity of a cube. *Nature* **350**, 631–633 (1991).
- Seeman, N. C. Nucleic acid nanostructures and topology. *Angew. Chem. Int. Edin Engl.* **37**, 3220–3238 (1998).
- Li, X., Yang, X., Qi, J. & Seeman, N. C. Antiparallel DNA double crossover molecules as components for nanoconstruction. *J. Am. Chem. Soc.* **118**, 6131–6140 (1996).
- Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394**, 539–544 (1998).
- Mao, C., Sun, W. & Seeman, N. C. Designed two-dimensional DNA Holliday junction arrays visualized by atomic force microscopy. *J. Am. Chem. Soc.* **121**, 5437–5443 (1999).
- LaBean, T. et al. The construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *J. Am. Chem. Soc.* **122**, 1848–1860 (2000).
- Mao, C., Sun, W., Shen, Z. & Seeman, N. C. A DNA nanomechanical device based on the B–Z transition. *Nature* **397**, 144–146 (1999).
- Yurke, B., Turberfield, A. J., Mills, A. P. Jr, Simmel, F. C. & Newmann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605–608 (2000).
- Yan, H., Zhang, X., Shen, Z. & Seeman, N. C. A robust DNA mechanical device controlled by hybridization topology. *Nature* **415**, 62–65 (2002).
- Niemeyer, C. M., Koehler, J. & Wuerdemann, C. DNA-directed assembly of bi-enzymic complexes from *in vivo* biotinylated NADP(H):FMN oxidoreductase and luciferase. *ChemBioChem* **3**, 242–245 (2002).
- Robinson, B. H. & Seeman, N. C. The design of a biochip: a self-assembling molecular-scale memory device. *Protein Eng.* **1**, 295–300 (1987).
- Keren, K. et al. Sequence-specific molecular lithography on single DNA molecules. *Science* **297**, 72–75 (2002).
- Alivisatos, A. P. et al. Organization of 'nanocrystal molecules' using DNA. *Nature* **382**, 609–611 (1996).
- Taton, T. A., Mucic, R. C., Mirkin, C. A. & Letsinger, R. L. The DNA-mediated formation of supramolecular mono- and multilayered nanoparticle structures. *J. Am. Chem. Soc.* **122**, 6305–6306 (2000).
- Pena, S. R. N., Raina, S., Goodrich, G. P., Fedoroff, N. V. & Keating, C. D. Hybridization and enzymatic extension of Au nanoparticle-bound oligonucleotides. *J. Am. Chem. Soc.* **124**, 7314–7323 (2002).
- Dekker, C. & Ratner, M. A. Electronic properties of DNA. *Phys. World* **14**, 29–33 (2001).
- Fahlman, R. P. & Sen, D. DNA conformational switches as sensitive electronic sensors of analytes. *J. Am. Chem. Soc.* **124**, 4610–4616 (2002).
- Seeman, N. C. The construction of 3-D stick figures from branched DNA. *DNA Cell Biol.* **10**, 475–486 (1991).
- Eckardt, L. H. et al. Chemical copying of connectivity. *Nature* **420**, 286 (2002).
- Adleman, L. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994).
- Winfree, E. in *DNA Based Computers. Proceedings of a DIMACS Workshop, April 4, 1995, Princeton University* (eds Lipton, R. J. & Baum, E. B.) 199–219 (American Mathematical Society, Providence, 1996).
- Winfree, E. Algorithmic self-assembly of DNA: theoretical motivations and 2D assembly experiments. *J. Biol. Mol. Struct. Dynamics Conversat.* **11**, 263–270 (2000).
- Mao, C., LaBean, T., Reif, J. H. & Seeman, N. C. Logical computation using algorithmic self-assembly of DNA triple crossover molecules. *Nature* **407**, 493–496 (2000).

Acknowledgements

This work has been supported by grants from the National Institute of General Medical Sciences, the Office of Naval Research, the National Science Foundation, and the Defense Advanced Research Projects Agency/Air Force Office of Scientific Research.

DNA replication and recombination

Bruce Alberts

National Academy of Sciences, 2101 Constitution Avenue,
Washington DC 20418, USA

Knowledge of the structure of DNA enabled scientists to undertake the difficult task of deciphering the detailed molecular mechanisms of two dynamic processes that are central to life: the copying of the genetic information by DNA replication, and its reassortment and repair by DNA recombination. Despite dramatic advances towards this goal over the past five decades, many challenges remain for the next generation of molecular biologists.

"Though facts are inherently less satisfying than the intellectual conclusions drawn from them, their importance should never be questioned."
James D. Watson, 2002.

DNA carries all of the genetic information for life. One enormously long DNA molecule forms each of the chromosomes of an organism, 23 of them in a human. The fundamental living unit is the single cell. A cell gives rise to many more cells through serial repetitions of a process known as cell division. Before each division, new copies must be made of each of the many molecules that form the cell, including the duplication of all DNA molecules. DNA replication is the name given to this duplication process, which enables an organism's genetic information — its genes — to be passed to the two daughter cells created when a cell divides. Only slightly less central to life is a process that requires dynamic DNA acrobatics, called homologous DNA recombination, which reshuffles the genes on chromosomes. In reactions closely linked to DNA replication, the recombination machinery also repairs damage that inevitably occurs to the long, fragile DNA molecules inside cells (see article in this issue by Friedberg, page 436).

The model for the DNA double helix¹ proposed by James Watson and Francis Crick is based on two paired DNA strands that are complementary in their nucleotide sequence. The model had striking implications for the processes of DNA replication and DNA recombination. Before 1953, there had been no meaningful way of even speculating about the molecular mechanisms of these two central genetic processes. But the proposal that each nucleotide in one strand of DNA was tightly base-paired with its complementary nucleotide on the opposite strand — either adenine (A) with thymine (T), or guanine (G) with cytosine (C) — meant that any part of the nucleotide sequence could act as a direct template for the corresponding portion of the other strand. As a result, any part of the sequence can be used either to create or to recognize its partner nucleotide sequence — the two functions that are central for DNA replication and DNA recombination, respectively.

In this review, I discuss how the discovery of the structure of DNA half a century ago opened new avenues for understanding the processes of DNA replication and recombination. I shall also emphasize how, as our understanding of complex biological molecules and their interactions increased over the years, there have been profound changes in the way that biologists view the chemistry of life.

Structural features of DNA

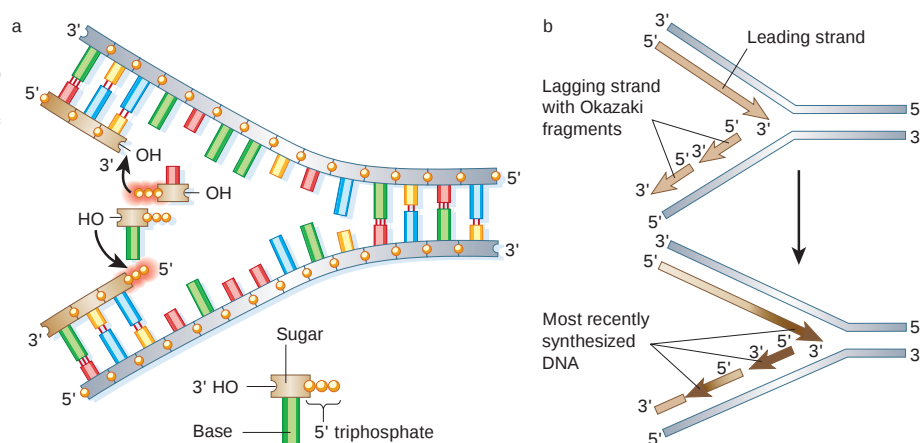
The research that immediately followed the discovery of the double helix focused primarily on understanding the structural properties

Figure 1 The DNA replication fork.

a. Nucleoside triphosphates serve as a substrate for DNA polymerase, according to the mechanism shown on the top strand.

Each nucleoside triphosphate is made up of three phosphates (represented here by yellow spheres), a deoxyribose sugar (beige rectangle) and one of four bases (differently coloured cylinders). The three phosphates are joined to each other by high-energy bonds, and the cleavage of these bonds during the polymerization reaction releases the free energy needed to drive the incorporation of each nucleotide into the growing DNA chain. The reaction shown on the bottom strand, which would cause DNA chain growth in the 3' to 5' chemical direction, does not occur in nature. **b.** DNA

polymerases catalyse chain growth only in the 5' to 3' chemical direction, but both new daughter strands grow at the fork, so a dilemma of the 1960s was how the bottom strand in this diagram was synthesized. The asymmetric nature of the replication fork was recognized by the early 1970s: the 'leading strand' grows continuously, whereas the 'lagging strand' is synthesized by a DNA polymerase through the backstitching mechanism illustrated. Thus, both strands are produced by DNA synthesis in the 5' to 3' direction. (Redrawn from ref. 27, with permission.)



of the molecule. DNA specifies RNA through the process of gene transcription, and the RNA molecules in turn specify all of the proteins of a cell. This is the 'central dogma' of genetic information transfer². Any read-out of genetic information — whether it be during DNA replication or gene transcription — requires access to the sequence of the bases buried in the interior of the double helix. DNA strand separation is therefore critical to DNA function. Thus, the Watson–Crick model drove scientists to a search for conditions that would disrupt the hydrogen bonds joining the complementary base pairs, so as to separate the two strands of the DNA double helix.

Physical chemists found that heating a solution of DNA to temperatures near boiling (100 °C), or subjecting it to extremes of pH, would cause the strands to separate — a change termed 'DNA denaturation'. The so-called 'melting temperature' (or T_m) of a stretch of DNA sequence depends on its nucleotide composition: those DNAs with a larger proportion of G–C base pairs exhibit a higher T_m because of the three hydrogen bonds that Watson and Crick had predicted to hold a G–C base pair together, compared with only two for the A–T base pair. At physiological salt concentrations, the T_m of mammalian DNA is nearly 90 °C, owing to the particular mix of its base pairs (47% G–C and 53% A–T)³.

Initially it seemed inconceivable that, once separated from its complementary partner, a DNA strand could reform a double helix again. In a complex mixture of DNA molecules, such a feat would require finding the one sequence match amongst millions during random collisions with other sequences, and then rapidly rewinding with a new partner strand. The dramatic discovery of this unexpected phenomenon⁴, called 'DNA renaturation', shed light on how sequences could be rearranged by DNA recombination. And it also provided a critical means by which DNA could be manipulated in the laboratory. The annealing of complementary nucleotide sequences, a process called hybridization, forms the basis of several DNA technologies that helped launch the biotechnology industry and modern genomics. These include gene cloning, genomic sequencing, and DNA copying by the polymerase chain reaction (see article by Hood and Galas on page 444).

The arrangement of DNA molecules in chromosomes presented another mystery for scientists: a long, thin molecule would be highly sensitive to shear-induced breakage, and it was hard to imagine that a mammalian chromosome might contain only a single DNA molecule. This would require that a typical chromosome be formed from a continuous DNA helix more than 100 million nucleotide pairs long

— a massive molecule weighing more than 100 billion daltons, with an end-to-end distance of more than 3 cm. How could such a giant molecule be protected from accidental fragmentation in a cell only microns in diameter, while keeping it organized for efficient gene readout and other genetic functions?

There was no precedent for such giant molecules outside the world of biology. But in the early 1960s, autoradiographic studies revealed that the chromosome of the bacterium *Escherichia coli* was in fact a single DNA molecule, more than 3 million nucleotide pairs in length⁵. And when — more than a decade later — innovative physical techniques demonstrated that a single huge DNA molecule formed the basis for each mammalian chromosome⁶, the result was welcomed by scientists with little surprise.

DNA replication forks

How is the enormously long double-stranded DNA molecule that forms a chromosome accurately copied to produce a second identical chromosome each time a cell divides? The template model for DNA replication, proposed by Watson and Crick in 1953 (ref. 7), gained universal acceptance after two discoveries in the late 1950s. One was an elegant experiment using density-labelled bacterial DNAs that confirmed the predicted template–anti-template scheme⁸. The other was the discovery of an enzyme called DNA polymerase, which uses one strand of DNA as a template to synthesize a new complementary strand⁹. Four deoxyribonucleoside triphosphate nucleotides — dATP, dTTP, dGTP and dCTP — are the precursors to a new daughter DNA strand, each nucleotide selected by pairing with its complementary nucleotide (T, A, C or G, respectively) on the parental template strand. The DNA polymerase was shown to use these triphosphates to add nucleotides one at a time to the 3' end of the newly synthesized DNA molecule, thereby catalysing DNA chain growth in the 5' to 3' chemical direction.

Although the synthesis of short stretches of DNA sequence on a single-stranded template could be demonstrated in a test tube, how an enormous, twisted double-stranded DNA molecule is replicated was a puzzle. Inside the cell, DNA replication was observed to occur at a Y-shaped structure, called a 'replication fork', which moves steadily along a parental DNA helix, spinning out two daughter DNA helices behind it (the two arms of the 'Y'). As predicted by Watson and Crick, the two strands of the double helix run in opposite chemical directions. Therefore, as a replication fork moves, DNA polymerase can move continuously along only one arm of the Y — the arm on

Box 1

Core proteins at the DNA replication fork

Proteins at the Y-shaped DNA replication fork are illustrated schematically in panel **a** of the figure below, but in reality, the fork is folded in three dimensions, producing a structure resembling that of the diagram in the inset **b** (cartoons redrawn from ref. 27, with permission).

Focusing on the schematic illustration in **a**, two DNA polymerase molecules are active at the fork at any one time. One moves continuously to produce the new daughter DNA molecule on the leading strand, whereas the other produces a long series of short 'Okazaki DNA fragments' on the lagging strand. Both polymerases are anchored to their template by polymerase accessory proteins, in the form of a sliding clamp and a clamp loader.

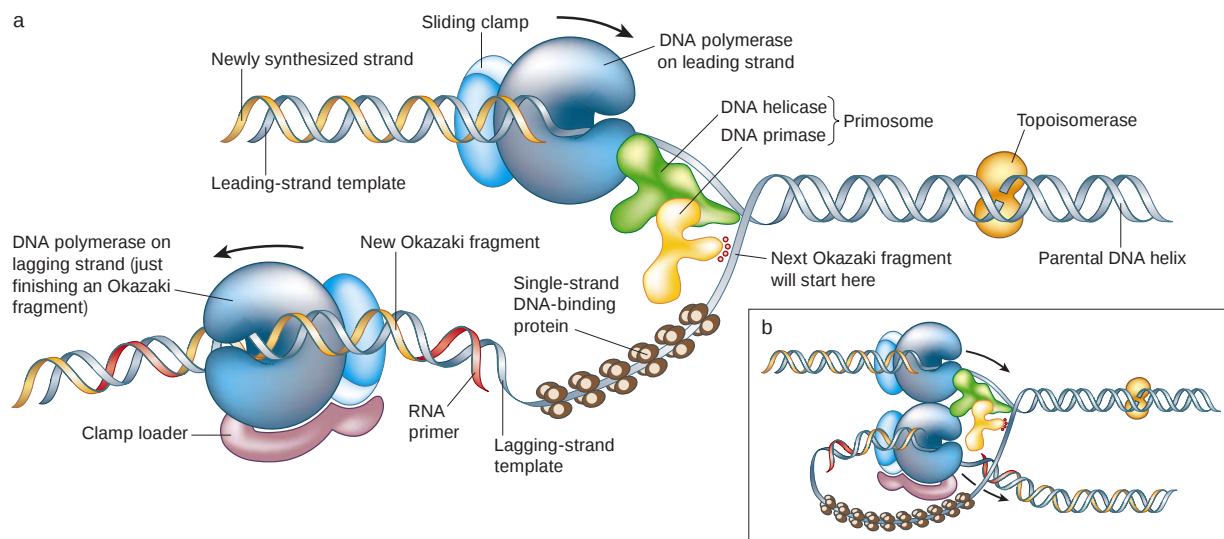
A DNA helicase, powered by ATP hydrolysis, propels itself rapidly along one of the template DNA strands (here the lagging strand), forcing open the DNA helix ahead of the replication fork. The helicase exposes the bases of the DNA helix for the leading-strand polymerase to copy. DNA topoisomerase enzymes facilitate DNA helix unwinding.

In addition to the template, DNA polymerases need a pre-existing DNA or RNA chain end (a primer) onto which to add each nucleotide. For this reason, the lagging strand polymerase requires the action of a

DNA primase enzyme before it can start each Okazaki fragment. The primase produces a very short RNA molecule (an RNA primer) at the 5' end of each Okazaki fragment onto which the DNA polymerase adds nucleotides. Finally, the single-stranded regions of DNA at the fork are covered by multiple copies of a single-strand DNA-binding protein, which hold the DNA template strands open with their bases exposed.

In the folded fork structure shown in the inset, the lagging-strand DNA polymerase remains tied to the leading-strand DNA polymerase. This allows the lagging-strand polymerase to remain at the fork after it finishes the synthesis of each Okazaki fragment. As a result, this polymerase can be used over and over again to synthesize the large number of Okazaki fragments that are needed to produce a new DNA chain on the lagging strand.

In addition to the above group of core proteins, other proteins (not shown) are needed for DNA replication. These include a set of initiator proteins to begin each new replication fork at a replication origin, an RNaseH enzyme to remove the RNA primers from the Okazaki fragments, and a DNA ligase to seal the adjacent Okazaki fragments together to form a continuous DNA strand.



which the new daughter strand is being elongated in the 5' to 3' chemical direction. On the other arm, the new daughter strand would need to be produced in the opposite, 3' to 5' chemical direction (Fig. 1a). So, whereas Watson and Crick's central predictions were confirmed at the end of the first decade of research that followed their landmark discovery, the details of the DNA replication process remained a mystery.

Reconstructing replication

The mystery was solved over the course of the next two decades, a period in which the proteins that constitute the central players in the DNA replication process were identified. Scientists used a variety of experimental approaches to identify an ever-growing set of gene products thought to be critical for DNA replication. For example, mutant organisms were identified in which DNA replication was defective, and genetic techniques could then be used to identify specific sets of genes required for the replication process^{10–12}. With the aid of the proteins specified by these genes, 'cell-free' systems were established, where the process was re-created *in vitro* using purified

components. Initially, proteins were tested in a 'partial replication reaction', where only a subset of the protein machinery required for the full replication process was present, and the DNA template was provided in a single-stranded form¹³. New proteins that were identified were added one at a time or in combination to test their effects on the catalytic activity of DNA polymerase. Further advances in understanding replication then depended on creating more complex *in vitro* systems, in which, through the addition of a larger set of purified proteins, double-stranded DNA could eventually be replicated^{14–15}.

Today, nearly every process inside cells—from DNA replication and recombination to membrane vesicle transport—is being studied in an *in vitro* system reconstructed from purified components. Although laborious to establish, such systems enable the precise control of both the concentration and the detailed structure of each component. Moreover, the 'noise' in the natural system caused by side reactions—because most molecules in a cell are engaged in more than one type of reaction—is avoided by eliminating the proteins that catalyse these other reactions. In essence, a small fraction of the cell can be re-created as a

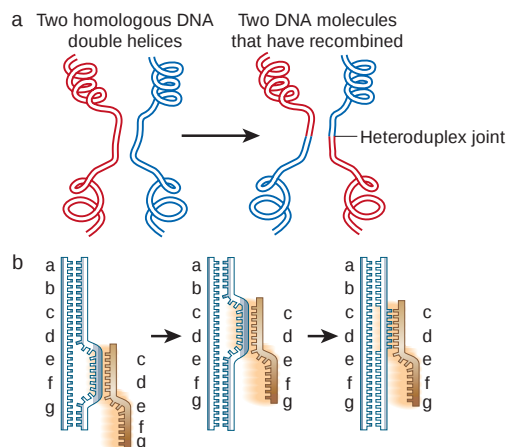
Box 2 DNA recombination

Homologous DNA recombination involves an exchange between two DNA double helices that causes a section of each helix to be exchanged with a section of the other, as illustrated schematically in panel **a** in the figure below (redrawn from ref. 27, with permission). Critical to the reaction is the formation of a heteroduplex joint at the point where the two double helices have been broken and then joined together. To form this joint, which glues two previously separate molecules together, a strand from one helix must form base pairs with a complementary strand from the second helix. This requires that the two DNA helices that recombine have a very similar sequence of nucleotides, that is, they must be homologous.

The DNA double helix poses a major problem for the DNA recombination process, because the bases that need to pair to form a heteroduplex joint are buried in the interior of the helix. How can two DNA helices recognize that they are homologous, in order to begin a recombination event, if their bases are not exposed?

The breakthrough came from the isolation and characterization of the RecA protein¹⁷ from the bacterium *Escherichia coli*, which would turn out to be the prototype for a family of strand-exchange proteins that is present in all organisms, from bacteria to humans. The human equivalent of the RecA protein is the Rad51 protein. These proteins catalyse the central synapsis step of homologous DNA recombination — the process that brings two matching DNA helices together and causes them to exchange parts, resulting in either the reassortment or the repair of genetic information (panel **b** below). Powered by the energy generated from ATP hydrolysis, the RecA protein assembles into long filaments on a single-strand DNA molecule (brown strand). Because the RecA protein has a second DNA-binding site that recognizes a DNA double helix, a RecA-coated strand has the remarkable ability to scan for a complementary strand in any double helix (blue strand) that it encounters. Once found, the complementary strand is pulled from the helix to form a new 'hybrid helix' with the RecA-coated single strand, thereby initiating the formation of the heteroduplex joint needed for recombination, as illustrated schematically in panel **b** (RecA protein not shown).

DNA recombination makes it possible for a damaged chromosome to repair itself by using a second copy of the same genetic information as a guide. It also causes the extensive breakage and reunion of chromosomes that occurs during the development of eggs and sperm, which greatly increases the genetic variation produced by sexual reproduction. Many of the atomic details of the RecA protein reaction are still uncertain, remaining as a future challenge for scientists.



bounded set of chemical reactions, making it fully amenable to precise study using all of the tools of physics and chemistry.

By 1980, multiprotein *in vitro* systems had enabled a detailed characterization of the replication machinery and solved the problem of how DNA is synthesized on both sides of the replication fork (Fig. 1b). One daughter DNA strand is synthesized continuously by a DNA polymerase molecule moving along the 'leading strand', while a second DNA polymerase molecule on the 'lagging strand' produces a long series of fragments (called Okazaki fragments)¹⁶ which are joined together by the enzyme DNA ligase to produce a continuous DNA strand. As might be expected, there is a difference in the proteins required for leading- and lagging-strand DNA synthesis (see Box 1). Remarkably, the replication forks formed in these artificial systems could be shown to move at the same rapid rates as the forks inside cells (500 to 1,000 nucleotides per second), and the DNA template was copied with incredibly high fidelity¹⁵.

As more and more proteins were found to function at the replication fork, comparisons could be made between the replication machinery of different organisms. Studies of the replication machinery in viruses, bacteria and eukaryotes revealed that a common set of protein activities drives the replication forks in each organism (Box 1). Each system consists of: a leading- and a lagging-strand DNA polymerase molecule; a DNA primase to produce the RNA primers that start each Okazaki fragment; single-strand DNA binding proteins that coat the template DNA and hold it in position; a DNA helicase that unwinds the double helix; and additional polymerase accessory proteins that tie the polymerases to each other and to the DNA template. As one progresses from a simple virus to more complex organisms, such as yeasts or mammals, the number of subunits that make up each type of protein activity tends to increase. For example, the total number of polypeptide subunits that form the core of the replication apparatus increases from four and seven in bacteriophages T7 and T4, respectively, to 13 in the bacterium *E. coli*. And it expands to at least 27 in the yeast *Saccharomyces cerevisiae* and in mammals. Thus, as organisms with larger genomes evolved, the replication machinery added new protein subunits, without any change in the basic mechanisms^{15,18–20}.

While the work I have described on DNA replication was advancing, other groups of researchers were establishing *in vitro* systems in which homologous DNA recombination could be reconstructed. The central player in these reactions was the RecA type of protein¹⁷, named after the bacterial mutant defective in recombination that led to its discovery (Box 2).

Protein machines

As for all other aspects of cell biochemistry, the DNA replication apparatus has evolved over billions of years through 'trial and error' — that is, by random variation followed by natural selection. With time, one protein after another could be added to the mix of proteins active at the replication fork, presumably because the new protein increased the speed, control or accuracy of the overall replication process. In addition, the structure of each protein was fine-tuned by mutations that altered its amino acid sequence so as to increase its effectiveness. The end results of this unusual engineering process are the replication systems that we observe today in different organisms. The mechanism of DNA replication might therefore be expected to be highly dependent on random past events. But did evolution select for whatever works, with no need for elegance?

For the first 30 years after Watson and Crick's discovery, most researchers seemed to hold the view that cell processes could be sloppy. This view was encouraged by knowledge of the tremendous speed of movements at the molecular level (for example, it was known that a typical protein collides with a second molecule present at a concentration of 1 mM about 10^6 times per second). The rapid rates of molecular movement were thought initially to allow a process like DNA replication to occur without any organization of the proteins involved in three-dimensional space.

Quite to the contrary, molecular biologists now recognize that evolution has selected for highly ordered systems. Thus, for example, not only are the parts of the replication machinery held together in precise alignments to optimize their mutual interactions, but energy-driven changes in protein conformations are used to generate coordinated movements. This ensures that each of the successive steps in a complex process like DNA replication is closely coordinated with the next one. The result is an assembly that can be viewed as a 'protein machine'. For example, the DNA polymerase molecule on the lagging side of the replication fork remains bound to the leading-strand DNA polymerase molecule to ensure that the same lagging-strand polymerase is used over and over again for efficient synthesis of Okazaki fragments^{18,20,21} (Box 1). And DNA replication is by no means unique. We now believe that nearly every biological process is catalysed by a set of ten or more spatially positioned, interacting proteins that undergo highly ordered movements in a machine-like assembly²².

Protein machines generally form at specific sites in response to particular signals, and this is particularly true for protein machines that act on DNA. The replication, repair and recombination of the DNA double helix are often considered as separate, isolated processes. But inside the cell, the same DNA molecule is able to undergo any one of these reactions. Moreover, specific combinations of the three types of reactions occur. For instance, DNA recombination is often linked directly to either DNA replication or DNA repair²³. For the integrity of a chromosome to be properly maintained, each specific reaction must be carefully directed and controlled. This requires that sets of proteins be assembled on the DNA and activated only where and when they are needed. Although much remains to be learned about how these choices are made, it seems that different types of DNA structures are recognized explicitly by specialized proteins that serve as 'assembly factors'. Each assembly factor then serves to nucleate a cooperative assembly of the set of proteins that forms a particular protein machine, as needed for catalysing a reaction appropriate to that time and place in the cell.

A view of the future

It has become customary, both in textbooks and in the regular scientific literature, to explain molecular mechanisms through simple two-dimensional drawings or 'cartoons'. Such drawings are useful for consolidating large amounts of data into a simple scheme, as illustrated in this review. But a whole generation of biologists may have become lulled into believing that the essence of a biological mechanism has been captured, and the entire problem therefore solved, once a researcher has deciphered enough of the puzzle to be able to draw a meaningful cartoon of this type.

In the past few years, it has become abundantly clear that much more will be demanded of scientists before we can claim to fully understand a process such as DNA replication or DNA recombination. Recent genome sequencing projects, protein-interaction mapping efforts and studies in cell signalling have revealed many more components and molecular interactions than were previously realized. For example, according to one recent analysis, *S. cerevisiae*, a single-celled 'simple' eukaryotic organism (which has about 6,000 genes compared with 30,000 in humans), uses 88 genes for its DNA replication and 49 genes for its DNA recombination²⁴.

To focus on DNA replication, fully understanding the mechanism will require returning to where the studies of DNA first began — in the realms of chemistry and physics. Detailed atomic structures of all relevant proteins and nucleic acids will be needed, and spectacular progress is being made by structural biologists, owing to increasingly powerful X-ray crystallography and nuclear magnetic resonance techniques. But the ability to reconstruct biological processes in a test tube with molecules whose precise structures are known is not enough. The replication process is both very rapid and incredibly accurate, achieving a final error rate of about one nucleotide in a billion. Understanding how the reactions between the many different proteins and

other molecules are coordinated to create this result will require that experimentalists determine all of the rate constants for the interactions between the various components, something that is rarely done by molecular biologists today. They can then use genetic engineering techniques to alter selected sets of these parameters, carefully monitoring the effect of these changes on the replication process.

Scientists will be able to claim that they truly understand a complex process such as DNA replication only when they can precisely predict the effect of changes in each of the various rate constants on the overall reaction. Because the range of experimental manipulations is enormous, we will need more powerful ways of deciding which such alterations are the most likely to increase our understanding. New approaches from the rapidly developing field of computational biology must therefore be developed — both to guide experimentation and to interpret the results.

The Watson–Crick model of DNA catalysed dramatic advances in our molecular understanding of biology. At the same time, its enormous success gave rise to the misleading view that many other complex aspects of biology might be similarly reduced to elegant simplicity through insightful theoretical analysis and model building. This view has been supplanted over subsequent decades, because most biological subsystems have turned out to be far too complex for their details to be predicted. We now know that nothing can substitute for rigorous experimental analyses. But traditional molecular and cell biology alone cannot bring a problem like DNA replication to closure. New types of approaches will be required, involving not only new computational tools, but also a greater integration of chemistry and physics^{20,25}. For this reason, we urgently need to rethink the education that we are providing to the next generation of biological scientists^{22,26}. □

doi:10.1038/nature01407

1. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738. (1953).
2. Crick, F. H. C. The biological replication of macromolecules. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
3. Doty, P. *Inside Nucleic Acids* (Harvey Lecture, 1960) (Academic, New York, 1961).
4. Marmur, J. & Doty, P. Thermal renaturation of deoxyribonucleic acids. *J. Mol. Biol.* **3**, 585–594 (1961).
5. Cairns, J. The bacterial chromosome and its manner of replication as seen by autoradiography. *J. Mol. Biol.* **6**, 208–213 (1963).
6. Kavenoff, R., Klotz, L. C. & Zimm, B. H. On the nature of chromosome-sized DNA molecules. *Cold Spring Harb. Symp. Quant. Biol.* **38**, 1–8 (1974).
7. Watson, J. D. & Crick, F. H. C. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
8. Meselson, M. & Stahl, F. W. The replication of DNA in *E. coli*. *Proc. Natl Acad. Sci. USA* **44**, 671–682 (1958).
9. Kornberg, A. Biological synthesis of DNA. *Science* **131**, 1503–1508 (1960).
10. Epstein, R. H. *et al.* Physiological studies of conditional lethal mutants of bacteriophage T4D. *Cold Spring Harb. Symp. Quant. Biol.* **28**, 375 (1963).
11. Bonhoeffer, F. & Schaller, H. A method for selective enrichment of mutants based on the high UV sensitivity of DNA containing 5-bromouracil. *Biochem. Biophys. Res. Commun.* **20**, 93 (1965).
12. Kohiyama, M., Cousin, D., Rytter, A. & Jacob, F. Mutants thermosensibles d'*Escherichia coli* K/12. I. Isolement et caractérisation rapide. *Ann. Inst. Pasteur* **110**, 465 (1966).
13. Huberman, J. A., Kornberg, A. & Alberts, B. M. Stimulation of T4 bacteriophage DNA polymerase by the protein product of T4 gene 32. *J. Mol. Biol.* **62**, 39–52 (1971).
14. Morris, C. F., Sinha, N. K. & Alberts, B. M. Reconstruction of bacteriophage T4 DNA replication apparatus from purified components: rolling circle replication following *de novo* chain initiation on a single-stranded circular DNA template. *Proc. Natl Acad. Sci. USA* **72**, 4800–4804 (1975).
15. Kornberg, A. & Baker, T. A. *DNA Replication* 2nd edn (Freeman, New York, 1992).
16. Okazaki R. *et al.* Mechanism of DNA chain growth: possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl Acad. Sci. USA* **59**, 598–605 (1968).
17. Radding, C. M. Recombination activities of *E. coli* RecA protein. *Cell* **25**, 3–4 (1981).
18. Davey, M. J. & O'Donnell, M. Mechanisms of DNA replication. *Curr. Opin. Chem. Biol.* **4**, 581–586 (2000).
19. Waga, S. & Stillman, B. The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* **67**, 721–751 (1998).
20. Benkovic, S. J., Valentine, A. M. & Salinas, F. Replisome-mediated DNA replication. *Annu. Rev. Biochem.* **70**, 181–208 (2001).
21. Alberts, B. M. The DNA enzymology of protein machines. *Cold Spring Harb. Symp. Quant. Biol.* **49**, 1–12 (1984).
22. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
23. Radding, C. Colloquium introduction. Links between recombination and replication: vital roles of recombination. *Proc. Natl Acad. Sci. USA* **98**, 8172 (2001).
24. Dwight, S. S. *et al.* *Saccharomyces Genome Database (SGD)* provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **30**, 69–72 (2002).
25. Trakselis, M. A. & Benkovic, S. J. Intrinsics in ATP-dependent clamp loading: variations across replication systems. *Structure* **9**, 999–1004 (2001).
26. National Research Council. *Bio2010: Undergraduate Education to Prepare Biomedical Research Scientists* (The National Academies Press, Washington DC, 2002).
27. Alberts, B. *et al.* *Molecular Biology of the Cell* 4th edn (Garland, New York, 2002).

DNA damage and repair

Errol C. Friedberg

Laboratory of Molecular Pathology, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9072, USA
(e-mail: friedberg.errol@pathology.swmed.edu)

The aesthetic appeal of the DNA double helix initially hindered notions of DNA mutation and repair, which would necessarily interfere with its pristine state. But it has since been recognized that DNA is subject to continuous damage and the cell has an arsenal of ways of responding to such injury. Although mutations or deficiencies in repair can have catastrophic consequences, causing a range of human diseases, mutations are nonetheless fundamental to life and evolution.

"We totally missed the possible role of ... [DNA] repair although ... I later came to realise that DNA is so precious that probably many distinct repair mechanisms would exist." Francis Crick, writing in *Nature*, 26 April 1974 (ref. 1).

This retrospective reflection by Francis Crick, penned two decades after he and James Watson reported the structure of DNA, hints at the early perception of DNA as a highly stable macromolecular entity. This prevailing view at the time significantly delayed serious consideration of biochemical processes such as mutation and repair. It was once suggested by Frank Stahl that "the possibility that ... genes were ... subject to the hurly-burly of both insult and clumsy efforts to reverse the insults, was unthinkable."²

But subsequent work on three 'R's' of DNA metabolism — replication (copying of DNA prior to each cell division), recombination (exchanges between different DNA molecules in a cell) and repair (restoration of altered DNA to its normal state) — revealed the dynamic state of DNA. It became apparent that DNA in all living organisms continually incurs a myriad of types of damage, and that cells have devised ingenious mechanisms for tolerating and repairing the damage. Failure of these mechanisms can lead to serious disease consequences, as well illustrated in the human hereditary diseases xeroderma pigmentosum (XP), hereditary non-polyposis colon cancer (HNPCC) and some forms of breast cancer. XP is characterized by about a 10,000-fold increased risk of skin cancer associated with sunlight exposure; individuals with HNPCC manifest an increased hereditary predisposition to colon (and other) cancer.

The roots of repair

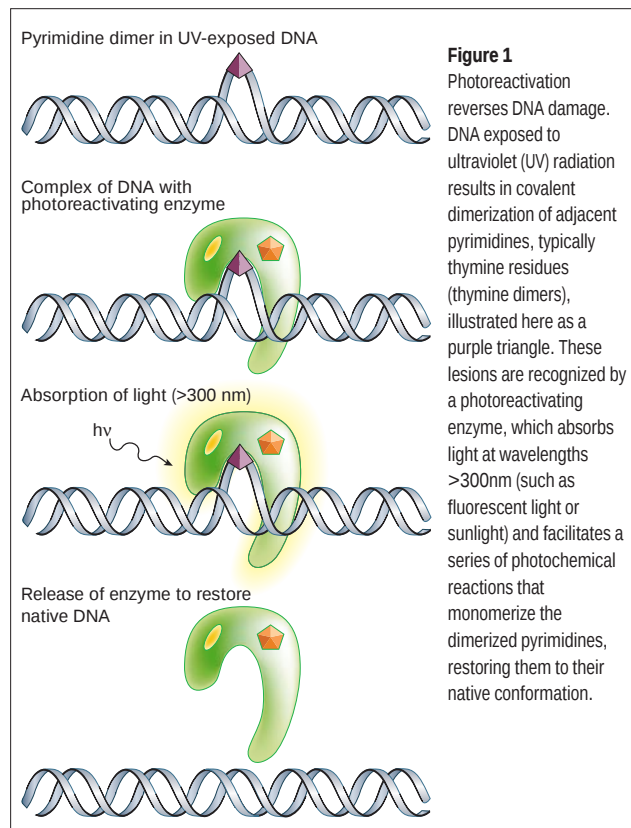
The early work on DNA damage and repair in the 1930s was stimulated by a small but prominent group of physicists³. As recounted by the geneticist Guido Pontecorvo, "in the years immediately preceding World War II something quite new happened: the introduction of ideas (not techniques) from the realm of physics into the realm of genetics, particularly as applied to the problems of size, mutability, and self-replication of genes"⁴. Seminal to this coalition between physics and biology in pre-war Germany was the collaboration between German physicists Karl Zimmer and Max Delbrück and the Russian geneticist Nikolai Timoféeff-Ressovsky⁵. Their partnership was stimulated by the work of Hermann Muller, a geneticist working

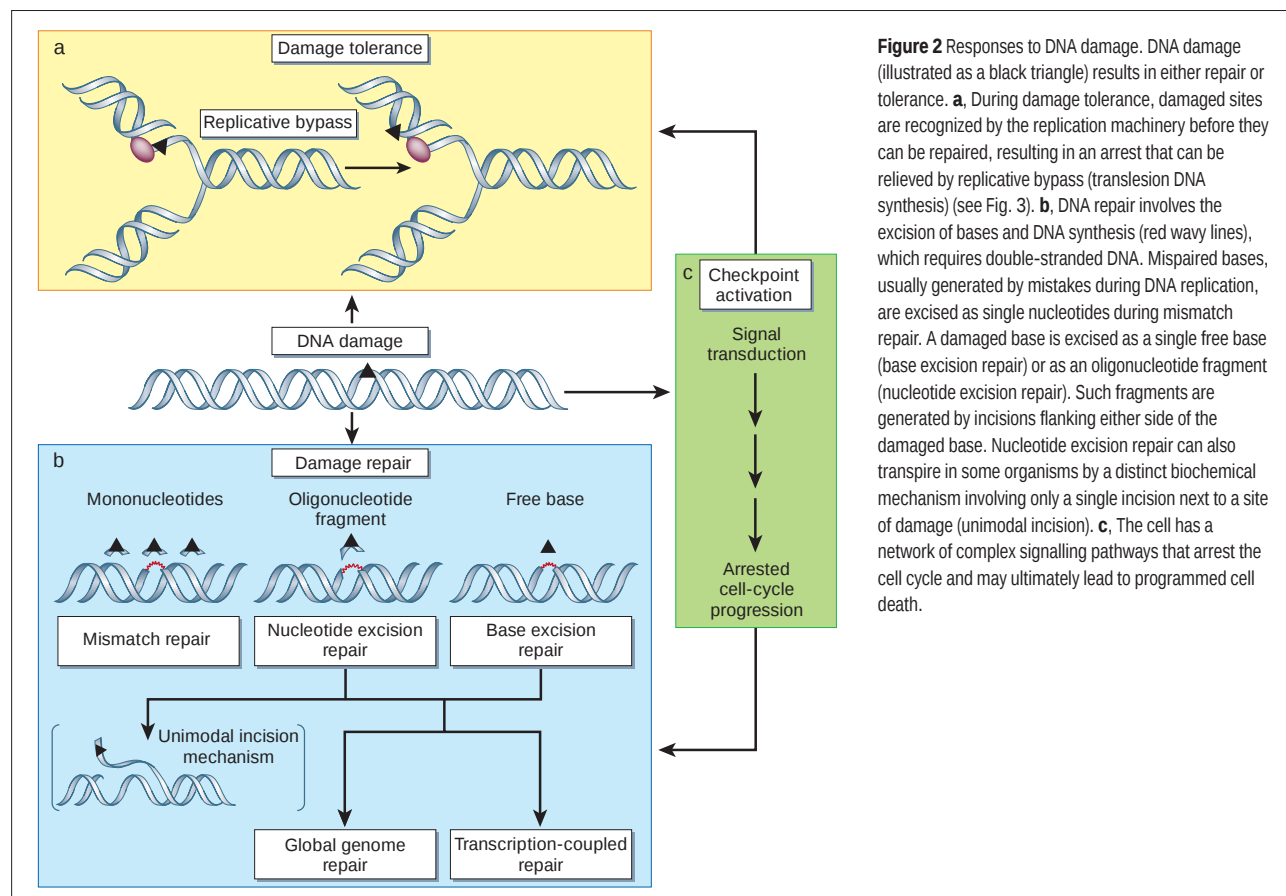
on the fruitfly *Drosophila* who first demonstrated that external agents, such as ionizing radiation, can cause mutations in living organisms⁶.

Timoféeff-Ressovsky and Zimmer were interested primarily in how such small amounts of energy in the form of ionizing radiation (formally equivalent to no more than the amount of energy absorbed as heat by drinking a cup of hot tea) could have such profound biological effects³. Delbrück and Muller, on the other hand, were intrigued by whether such mutations could reveal insight into the physical nature of the gene.

In retrospect, it was inevitable that the deployment of physical (and later chemical) tools, such as ionizing and ultraviolet (UV) radiation, to study genes would in due course lead to questions as to how these agents damaged DNA³. And, once it was recognized that these interactions promoted deleterious effects on the structure and function of genes, to questions concerning how cells cope with damaged DNA. Zimmer wrote, "one cannot use radiations for elucidating the normal state of affairs without considering the mechanisms of their actions, nor can one find out much about radiation induced changes without being interested in the normal state of the material under investigation."⁷

Hints of the ability of living cells to recover from the lethal effect of UV radiation emerged as early as the mid-1930s⁸. But the discovery of a DNA-repair mechanism had to wait until the end of the 1940s, through the independent, serendipitous observations of Albert Kelner⁹ working in Milislav Demerec's group at the Cold Spring Harbor Laboratory, and Renato Dulbecco¹⁰ in Salvador Luria's laboratory at The University of Indiana. Neither Kelner nor Dulbecco set out to study damage to DNA or its repair. They were both using UV radiation as an experimental tool, but observed anomalous survival rates when cells or bacteriophage (bacteria-infecting viruses) were inadvertently exposed to long-wavelength light, either as sunlight or fluorescent light in their respective laboratories^{9,10}. Their efforts to explain these confounding observations led to the discovery of the phenomenon now known as photoreaction-





vation, whereby the DNA damage incurred by exposure to UV light is repaired by a light-dependent enzyme reaction¹¹ (Fig. 1).

Curiously, even with the elucidation of the structure of DNA only four years away, neither Dulbecco nor Watson — who was a graduate student in Luria's laboratory when Dulbecco stumbled on photoreactivation, and had himself examined the effects of ionizing radiation for his doctoral thesis² — thought about DNA repair. However, shortly after Watson and Crick reported on the DNA double helical structure, they noted the implications of the base-pairing rules for mutagenesis, stating “spontaneous mutation may be due to a base occasionally occurring in one of its less likely tautomeric forms”¹².

Tautomerism is the property of a compound that allows it to exist in two interconvertible chemical states; in the case of DNA bases, as either keto or enol forms. Watson and Crick had initially overlooked the complications of tautomerism and were trying unsuccessfully to construct their DNA model with the rare enol form of bases. It was only after Jerry Donohue, a former graduate student of Linus Pauling, pointed out to Watson that he should be using the more common keto form that the problem of how bases could stably pair was solved¹³.

But no consideration was then given to the fact that the chemical lability of DNA implicit in tautomerism might have wider implications for the stability of genes. Indeed, the field gave little thought to the precise nature of DNA damage and its possible biological consequences. One must recall, however, that even at the time the DNA double helix was unveiled, its ‘pathology’ and the biological consequences thereof were far less compelling problems than deciphering the genetic code or understanding the essential features of DNA replication. Even mutagenesis — put to extensive use as a tool for determining the function of genes and their polypeptide products, and for defining the genetic code — was not widely considered in mechanistic terms until much later¹¹. This is despite the fact that

the repair phenomenon of photoreactivation was known before the discovery of the structure of DNA.

A DNA duplex for redundancy

Watson and Crick noted, with infamous prophetic understatement, “it has not escaped our notice that the specific [base] pairing we have postulated immediately suggests a possible copying mechanism for the genetic material”¹⁴. However, it was not intuitively obvious that a double-stranded molecule should be required for DNA replication. In principle, a single-stranded chain could just as easily do. But the significance of the duplex DNA structure soon became apparent. It was shown that DNA replicates in a semi-conservative fashion, whereby each strand of the double helix pairs with a new strand generated by replication. This enables errors introduced during DNA replication to be corrected by a mechanism known as excision repair, which relies on the redundancy inherent in having two complementary strands of the genetic code. If the nucleotides on one strand are damaged, they can be excised and the intact opposite strand used as a template to direct repair synthesis of DNA¹⁵ (Fig. 2).

Many paths to mutation and repair

The elucidation of the DNA structure provided the essential foundation for defining the different types of mutations arising from both spontaneous and environmental DNA damage that affect all living cells¹². Once again, the insights of physicists featured prominently³, including among others, Richard Setlow who identified thymine dimers as stable and naturally occurring DNA lesions arising in cells exposed to sunlight (UV radiation). Such lesions comprise a covalent joining of two adjacent thymine residues in the same DNA chain. They generate considerable distortion of the normal structure of DNA and seriously impede DNA transactions such as replication and transcription. The repair of these lesions could be monitored

experimentally, and promoted the discovery by Setlow¹⁶ and others² of excision repair in bacteria and higher organisms¹⁷.

As the profusion of alterations in DNA became more widely recognized, scientists came to appreciate that the identification of any new type of naturally occurring base damage would, if one searched diligently enough, almost certainly lead to the discovery of one or more mechanisms for its repair or tolerance^{2,18}. Such has indeed been the case. DNA repair now embraces not only the direct reversal of some types of damage (such as the enzymatic photoreactivation of thymine dimers), but also multiple distinct mechanisms for excising damaged bases, termed nucleotide excision repair (NER), base excision repair (BER) and mismatch repair (MMR)¹¹ (Fig. 2). The principle of all three mechanisms of repair involves splicing out the damaged region and inserting new bases to fill the gap, followed by ligation of the pieces.

The process of NER is biochemically complicated, involving as many as 30 distinct proteins in human cells that function as a large complex called the nucleotide excision repairosome. This 'repair machine' facilitates the excision of damaged nucleotides by generating bimodal incisions in the flanking regions and removing a fragment about 30 nucleotides in length¹¹ (Fig. 2). Damaged bases that are not recognized by the NER machinery are corrected by BER, whereby the bases are excised from the genome as free bases by a different set of repair enzymes. In MMR, incorrect bases incorporated as a result of mistakes during DNA replication are excised as single nucleotides by yet a third group of repair proteins (Fig. 2). Both NER and BER transpire by somewhat different mechanisms depending on whether the DNA damage is located in regions of the genome that are undergoing active gene expression (transcription-coupled repair) or are transcriptionally silent (global genome repair)^{11,19}.

In addition to the various modes of excision repair that evolved to cope with damaged bases or mistakes during replication, cells frequently suffer breakage of one or both chains of the DNA duplex¹¹. Naturally occurring reactive oxygen molecules and ionizing radiation are prevalent sources of such damage¹¹. Strand breaks must be repaired in order to maintain genomic integrity. In particular, double-strand breaks (DSBs) sever the chromosomes and are lethal unless repaired¹¹.

Several mechanisms for the repair of DSBs have been elucidated (Fig. 3). One of these involves swapping equivalent regions of DNA

between homologous chromosomes — a process called recombination¹¹. This type of exchange occurs naturally during meiosis, the special type of cell division that generates the germ cells (sperm and ova). It can also be used to repair a damaged site on a DNA strand by using information located on the undamaged homologous chromosome. This process requires an extensive region of sequence homology between the damaged and template strands. Multiple proteins are required for DSB repair by recombination and deficiencies in this repair mechanism can cause cancer. For example, mutation of at least one of these repair proteins (called BRCA1) causes hereditary breast cancer. An alternative mechanism for the repair of DSBs, called non-homologous end joining, also requires a multi-protein complex, and essentially joins broken chromosome ends in a manner that does not depend on sequence homology and may not be error free (Fig. 3).

Damage tolerance

Although insights into DNA repair have progressed at an impressive pace, especially in the past decade, an understanding of the mechanisms of mutagenesis — a phenomenon that, as mentioned earlier, was demonstrated experimentally before discovery of the structure of DNA — has lagged. A breakthrough came from the experimental demonstration that some mutations arise as a consequence of a cell's efforts to tolerate damage. In this situation, the base damage and/or strand breaks in DNA persist in the genome, but their potential for interfering with DNA replication and transcription is somehow mitigated.

One such damage-tolerance mechanism, called translesion DNA synthesis, involves the replication machinery bypassing sites of base damage, allowing normal DNA replication and gene expression to proceed downstream of the (unrepaired) damage²⁰ (Fig. 4). It involves specialized low-fidelity ('sloppy') DNA polymerases that are able to bypass DNA lesions that typically stall the high-fidelity polymerases required for DNA replication. To overcome the block, these 'sloppy copiers' add nucleotides to the replicating strand opposing the DNA lesion, thus allowing replication to continue, but nevertheless introducing mutations into the newly synthesized sequence²⁰.

Cell suicide

Recent years have witnessed the recognition that biological responsiveness to genetic insult embraces more than the repair and

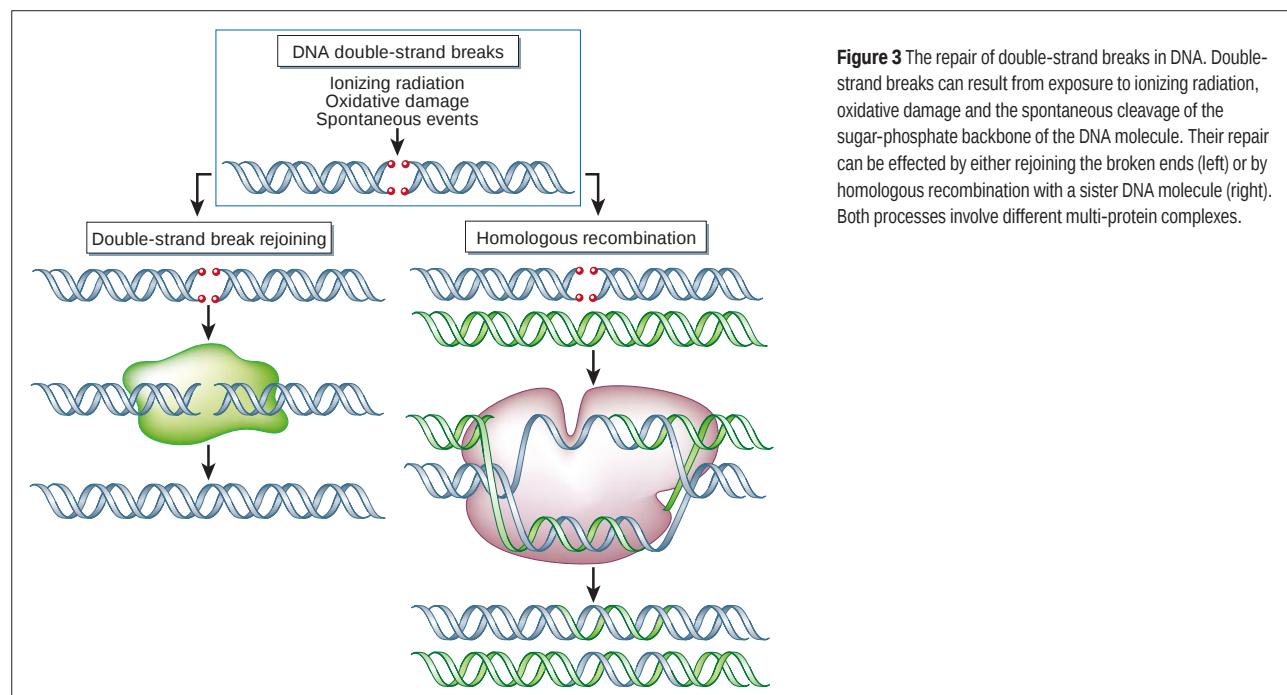


Figure 3 The repair of double-strand breaks in DNA. Double-strand breaks can result from exposure to ionizing radiation, oxidative damage and the spontaneous cleavage of the sugar-phosphate backbone of the DNA molecule. Their repair can be effected by either rejoining the broken ends (left) or by homologous recombination with a sister DNA molecule (right). Both processes involve different multi-protein complexes.

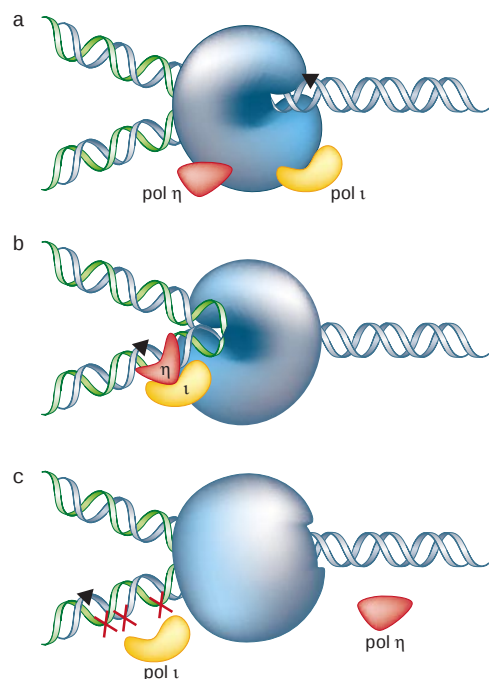


Figure 4 ‘Sloppy copiers’ overcome blocks in replication caused by a DNA lesion (a process called translesion synthesis). **a**, The DNA replicative machinery (blue) stalls immediately behind a site of base damage (black triangle). Two specialized ‘sloppy copier’ polymerases ($\text{pol } \eta$ and $\text{pol } \iota$) bind to the arrested replication complex. **b**, This interaction promotes a conformational change in the arrested replication machinery, placing $\text{pol } \eta$ in direct proximity to the site of base damage where it synthesizes across the lesion. **c**, $\text{Pol } \eta$ may then dissociate and allow $\text{pol } \iota$ to complete the process of replicative bypass by incorporating several more nucleotides (red crosses). Once the lesion has been completely bypassed, the replication machinery resumes DNA replication. As a result of this process, mutations to the DNA sequence are now incorporated into one strand.

tolerance of DNA damage. The exposure of cells to many DNA-damaging agents results in the transcriptional upregulation of a large number of genes, the precise function(s) of many of which remains to be established. Additionally, cells have evolved complex signalling pathways to arrest the progression of the cell cycle in the presence of DNA damage, thereby providing increased time for repair and tolerance mechanisms to operate²¹ (Fig. 2c). Finally, when the burden of genomic insult is simply too large to be effectively met by the various responses discussed, cells are able to initiate programmed cell death (apoptosis), thereby eliminating themselves from a population that otherwise might suffer serious pathological consequences²².

DNA damage and cancer

The ‘somatic mutation hypothesis’ of cancer embraces the notion that neoplastic transformation arises from mutations that alter the function of specific genes (now called oncogenes and tumour-suppressor genes) that are critical for cell division. This theory has its roots in correlations between chromosomal abnormalities and cancer first observed by the developmental biologist Theodore Boveri²³, who at the beginning of the twentieth century reported abnormal numbers of chromosomes (aneuploidy) in cancerous somatic cells.

The discovery of the structure of DNA progressed our understanding of tumorigenesis at several levels. Watson and Crick predicted from their DNA model that complementary base pairing had implications for recombination (the exchange of genetic

material between chromosome pairs): “the pairing between homologous chromosomes at meiosis may depend on pairing between specific bases”. The genetic basis of many cancers is now known to arise from abnormal recombination events, such as chromosomal translocations, where a region of one chromosome is juxtaposed to another chromosome. Watson himself developed an early and ardent interest in cancer biology when he recognized that the experimentally tractable genomes of oncogenic viruses could provide important insights into the pathogenesis of cancer. Mutagenesis is now documented as a fundamental cornerstone of the molecular basis of all forms of cancer²⁴.

Arguably the most definitive validation of the somatic mutation hypothesis derives from the discovery that defective responses to DNA damage and the accumulation of mutations underlies two distinct types of hereditary cancer; skin cancer associated with defective NER and colorectal cancer associated with defective MMR¹¹. In both instances, credit belongs to scholars of DNA repair.

In the late 1960s, James Cleaver providentially noted an article in the *San Francisco Chronicle* that reported the extreme proneness to skin cancer in individuals with XP, a rare sun-sensitive hereditary disease². Cleaver was then searching for mammalian cell lines that were defective in excision repair, and his intuitive notion that XP individuals might be sunlight-sensitive and prone to cancer because they were genetically defective in excision repair proved to be correct²⁵.

The subsequent elucidation of the genes defective in XP patients²⁶, and their role in NER of damaged bases in human cells^{11,27,28}, represents a triumph of modern genetics and its application to molecular biology. The additional discovery that the process of NER in eukaryotes requires elements of the basic transcription apparatus¹¹ has yielded insights into the complex relationships between deficient DNA repair, defective transcription and hereditary human diseases¹¹.

A fascinating denouement to the skin-cancer predisposition in XP patients derives from the recent solution of the ‘XP variant problem’. A significant fraction of XP individuals who are clinically indistinguishable from those defective in NER were found to be proficient in this repair process¹¹. It was shown that DNA polymerase- η ($\text{pol } \eta$), one of the specialized DNA polymerases capable of overcoming replication blocks at DNA lesions, is mutated in all XP-variant patients so far examined²⁹. Not only does $\text{pol } \eta$ replicate past thymine dimers in DNA, but — unlike the other specialized DNA polymerases — it also correctly inserts adenine residues²⁹, thereby preventing mutations at sites of thymine dimers. Therefore, even in XP patients with functional NER, in the absence of $\text{pol } \eta$ one or more other bypass polymerases attempts to cope with arrested replication at thymine dimers, but does so inaccurately²⁹. Thus, cancer predisposition in XP essentially derives from an excessive mutational burden in skin cells associated with exposure to sunlight. These mutations accumulate either because thymine dimers are not excised (owing to defective NER) or because in the absence of $\text{pol } \eta$, dimers are inaccurately bypassed by other DNA polymerases²⁹.

The association between HNPCC and defective MMR was determined more-or-less simultaneously by a number of investigators. Paul Modrich² surmised that the instability of repeated sequences in DNA associated with defective MMR in bacteria³⁰ might be causally related to the DNA sequence instability observed in patients with HNPCC³¹. This led to the formal demonstration of defective MMR in this human hereditary disease and formed another persuasive validation of the somatic mutation theory of cancer^{32–34}.

A look to the future

The study of biological responsiveness to DNA damage embraces DNA repair, mutagenesis, damage tolerance, cell-cycle checkpoint control, programmed cell death, and other cellular responses to

genomic insult. This integrated field is now deciphering the complex regulatory pathways transduced by signalling mechanisms that detect DNA damage and/or arrested DNA replication. As these pathways become better understood, parallel technological gains in gene therapy and therapeutic intervention by rational drug design will offer new strategies for blocking the unwanted consequences of DNA damage, especially cancer.

We must remember, however, that while evolution could not have transpired without robust cellular mechanisms to ameliorate the most serious consequences of spontaneous and environmental DNA damage, the process of evolution mandates that the genetic diversification on which Darwinian selection operates be maintained constantly. Thus, life is necessarily a delicate balance between genomic stability and instability — and of mutation and repair. □

doi:10.1038/nature01408

1. Crick, F. The double helix: a personal view. *Nature* **248**, 766–769 (1974).
2. Friedberg, E. C. *Correcting the Blueprint of Life: An Historical Account of the Discovery of DNA Repair Mechanisms* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1997).
3. Friedberg, E. C. The intersection between the birth of molecular biology and the DNA repair and mutagenesis field. *DNA Repair* **1**, 855–867 (2002).
4. Pontecorvo, G. *Trends in Genetic Analysis* (Columbia Univ. Press, New York, 1958).
5. Timoféeff-Ressovsky, N. W., Zimmer, K. G. & Delbrück, M. Über die Natur der Genmutation und der Genstruktur. *Nachr. Ges. Wiss. Göttingen FG VI Biol. N.F.* **1**, 189–245 (1935).
6. Muller, H. J. Artificial transmutation of the gene. *Science* **66**, 84–87 (1927).
7. Zimmer, K. G. in *Phage and the Origins of Molecular Biology* Expanded edn (eds Cairns, J., Stent, G. S. & Watson, J. D.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992).
8. Hollaender, A. & Curtis, J. T. Effect of sublethal doses of monochromatic ultraviolet radiation on bacteria in liquid suspension. *Proc. Soc. Exp. Biol. Med.* **33**, 61–62 (1935–36).
9. Kellner, A. Effect of visible light on the recovery of *Streptomyces griseus conidia* from ultraviolet irradiation injury. *Proc. Natl Acad. Sci. USA* **35**, 73–79 (1949).
10. Dulbecco, R. Reactivation of ultra-violet-inactivated bacteriophage by visible light. *Nature* **162**, 949–950 (1949).
11. Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (American Society of Microbiology Press, Washington DC, 1995).
12. Watson, J. D. & Crick, F. H. C. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
13. Watson, J. D. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (Atheneum, New York, 1968).
14. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
15. Hanawalt, P. C. & Haynes, R. H. The repair of DNA. *Sci. Am.* **216**, 36–43 (1967).
16. Setlow, R. B. & Carrier, W. L. The disappearance of thymine dimers from DNA: an error-correcting mechanism. *Proc. Natl Acad. Sci. USA* **51**, 226–231 (1964).
17. Rasmussen, D. E. & Painter, R. B. Evidence for repair of ultra-violet damaged deoxyribonucleic acid in cultured mammalian cells. *Nature* **203**, 1360–1362 (1964).
18. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
19. Le Page, F. *et al.* Transcription-coupled repair of 8-oxoguanine: requirement for XPG, TFIIH, and CSB and implications for Cockayne syndrome. *Cell* **101**, 159–171 (2000).
20. Goodman, M. F. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu. Rev. Biochem.* **71**, 17–50 (2002).
21. Zhou, B.-B. S. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).
22. Cory, S. & Adams J. M. The Bcl2 family: regulators of the cellular life-or-death switch. *Nature Rev. Cancer* **2**, 647–656 (2002).
23. Boveri, T. *The Origin of Malignant Tumors* (Williams and Wilkins, Baltimore, MD, 1929).
24. Vogelstein, B. & Kinzler, K. W. (eds) *The Genetic Basis of Human Cancer* (McGraw-Hill, New York, 1998).
25. Cleaver, J. E. Defective repair replication of DNA in xeroderma pigmentosum. *Nature* **218**, 652–656 (1968).
26. De Weerd-Kastelein, E. A., Keijzer, W. & Bootsma, D. Genetic heterogeneity of xeroderma pigmentosum demonstrated by somatic cell hybridization. *Nature New Biol.* **238**, 80–83 (1972).
27. Wood, R. D. DNA repair in eukaryotes. *Annu. Rev. Biochem.* **65**, 135–167 (1996).
28. Sancar, A. DNA excision repair. *Annu. Rev. Biochem.* **65**, 43–81 (1996).
29. Masutani, C. *et al.* Xeroderma pigmentosum variant: from a human genetic disorder to a novel DNA polymerase. *Cold Spring Harbor Symp. Quant. Biol.* **65**, 71–80 (2000).
30. Levinson, G., & Gutman, G. A. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res.* **15**, 5323–5338 (1987).
31. Aaltonen, L. A. *et al.* Clues to the pathogenesis of familial colorectal cancer. *Science* **260**, 812–816 (1993).
32. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
33. Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215–1225 (1993).
34. Parsons, R. *et al.* Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75**, 1227–1236 (1993).

Acknowledgements

I thank numerous individuals for their constructive comments. This article is dedicated to James Watson with thanks for his inspiration as a writer and historian.

The double helix and immunology

Gustav J. V. Nossal

Department of Pathology, The University of Melbourne, Victoria 3010, Australia

The immune system can recognize and produce antibodies to virtually any molecule in the Universe. This enormous diversity arises from the ingenious reshuffling of DNA sequences encoding components of the immune system. Immunology is an example of a field completely transformed during the past 50 years by the discovery of the structure of DNA and the emergence of DNA technologies that followed.

“This short history of research in one area, lymphocyte receptors, is yet another witness to the power of DNA technology, and to the ability of this approach not only to explain known biological phenomena, but also to contribute to the discovery of new biological systems.”

Susumu Tonegawa, Nobel lecture, 8 December 1987.

The double helix is all about biological information: how it is encoded, stored, replicated and used when required. Immunology, too, is about information. What genetic processes control the vast array of synthetic potential within an immune system capable of reacting specifically to virtually any microbe or foreign molecule? The secret lies in unique DNA processing that occurs during the development of lymphocyte cells, which are responsible for the specific immune response to a foreign agent (antigen). B lymphocytes produce antibodies (on their surface as well as secreted) and T lymphocytes mount cellular attacks on pathogenic infiltrators.

As lymphocytes develop, an array of short genes are rearranged and assembled together at the DNA level to form genes whose products recognize distinct antigens. As the process is mostly random, each lymphocyte makes different choices and thus the result is a vast repertoire of lymphocytes reactive to different antigens. This process has implications for antibody formation, cell-mediated immunity and malignancies of the immune system.

One B cell produces one antibody

At the beginning of the last century, Paul Ehrlich¹ recognized that the specificity of antibodies lay in the complementarity of their shapes to the antigen(s) on the microbe being recognized. He saw antibodies as cellular ‘side chains’, which budded out from the cell surface as what today we would term receptors. Karl Landsteiner² then demonstrated the exquisite specificity of antibodies, showing that animals could make antibodies to almost anything, including small synthetic organic molecules that had never previously existed in nature. Moreover, tiny structural changes in the antigen could lead to the production of a different antibody. It beggared belief that there could be so many different side chains. When antibodies were shown to be proteins, it seemed natural to conclude that a specific antibody molecule was shaped in close proximity to an antigen molecule much as plastic or sheet metal is moulded against a template. This ‘direct template’ hypothesis³ held sway for several decades.

In 1955, Niels Jerne⁴ published his natural selection theory of antibody formation, which postulated the random synthesis of a million or more different sorts of antibodies. When an antigen enters the body, it unites with an antibody that just happens to fit it, the antigen–antibody complex is taken up by a cell and the antibody

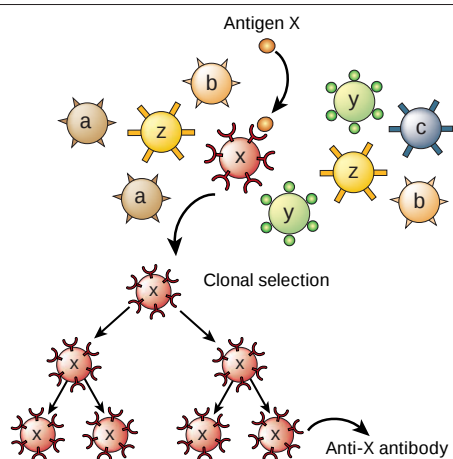


Figure 1 The clonal selection theory of antibody formation. Each B lymphocyte produces only one type of antibody receptor on its surface. An antigen recognizes one B lymphocyte out of a large repertoire. This triggers the rapid division and differentiation of the B cell to become a 'plasma cell', producing and secreting antibodies specific to the original antigen.

somehow acts as the template for the formation of more of itself. David Talmage⁵ and Macfarlane Burnet⁶ recognized that this theory would make more sense if the postulated natural antibodies were located on the surface of what we now call B lymphocyte cells. If each cell were endowed with only one sort of antibody specificity, then the antigen could select one lymphocyte out of a repertoire, cause its clonal division and stimulate antibody production and secretion (Fig. 1). In 1958, Joshua Lederberg⁷ and I provided the first evidence for the clonal selection theory, namely that one B cell always produces only one antibody.

DNA shuffling in antibody formation

Antibodies are multichain proteins that come in different forms. The most abundant, immunoglobulin- γ (IgG), consists of two identical light (L) chains and two identical heavy (H) chains⁸ (Fig. 2a). The

carboxy-terminal halves of the two light chains are identical to each other, but the amino-terminal halves differ in more than 50 residues, called the 'variable' region⁹. The heavy chains, too, consist of a variable (V) part and a constant (C) part.

In 1965, William J. Dreyer and J. Claude Bennett¹⁰ bucked the dogma at the time that 'one gene makes one protein', and put forward the revolutionary concept that the carboxy-terminal C region of the L chain was always encoded by a single gene, but that the amino-terminal V half could be encoded by multiple separate genes, perhaps as many as 100,000 in number. It followed that a chosen V gene must then somehow become associated with a C gene by a DNA rearrangement event in each lymphocyte cell, because only when the V–C regions were spliced together could a functional protein be expressed.

At that time, there was no way to interrogate the genome directly to test this concept, and for a decade debate and controversy raged. One school, led by Leroy Hood, favoured the idea that a large array of germline-encoded V genes for the L and H chains underwent rearrangement. At the other extreme were proponents of a single, very highly mutable V gene that was extensively mutated in emerging B cells. In the middle were those who favoured the idea of a handful of V genes that were subject to extensive recombination in somatic cells (the cells of the body, excluding the sex cells). This compromise was supported by luminaries such as Oliver Smithies and Gerald Edelman. Francis Crick was quite taken with the idea that just two V genes undergo rearrangement in the germ line, with further mutation in somatic cells.

Before arriving at the solution brought by advances in molecular biology, one more fact is worthy of note. Elvin Kabat astutely pointed out that in the V regions of both heavy and light chains there were also three short stretches of amino acids where variation was considerably greater than elsewhere in the molecules, and these so-called hypervariable regions were deemed likely to be sites of union with the antigen¹¹. Could it be that there was actually an assembly of several, rather than just two, genes encoding each chain?

The new tools for manipulating and sequencing DNA came to the rescue. In 1976, Nobumichi Hozumi and Susumu Tonegawa¹² conducted a landmark experiment. They used a DNA cutting enzyme known as a restriction endonuclease to digest the DNA extracted from a mouse embryo and from an antibody-secreting tumour. The resulting DNA fragments were then separated on the basis of size, and

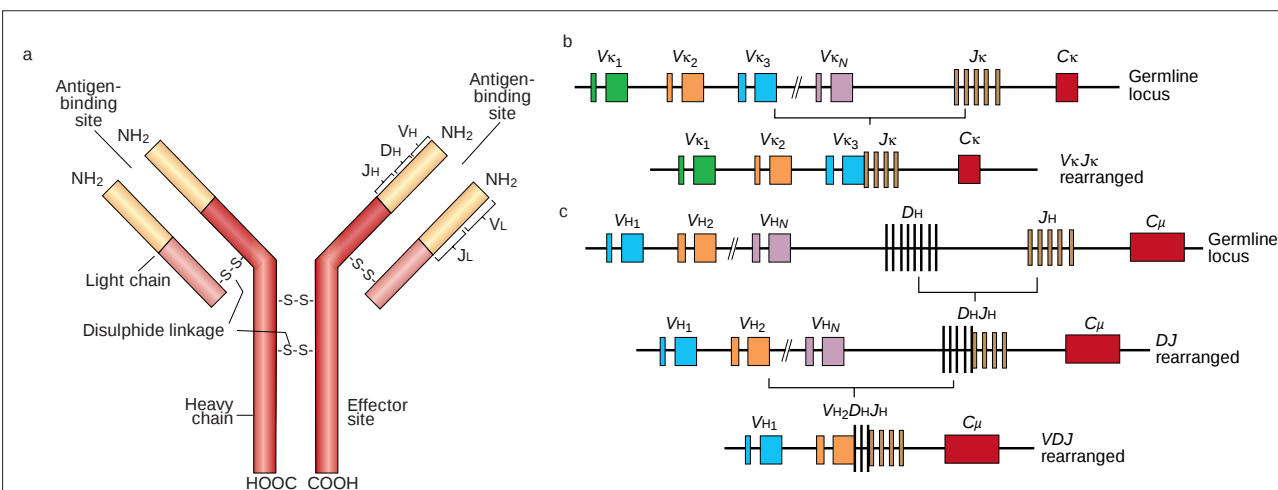


Figure 2 Antibody formation. **a**, Structure of an antibody (immunoglobulin). Two identical heavy chains are connected by disulphide linkages. The antigen-recognizing site is composed of the variable regions (yellow) of the heavy and light chains, whereas the effector site (which determines its function) is determined by the amino-acid sequence of the heavy chain constant region (red). **b**, Assembly of the light (κ)- and heavy (H)-chain genes of antibodies by somatic recombination during B-lymphocyte development. The L

chain is encoded by variable (V), joining (J) and constant (C) genes. While the developing B cell is still maturing in the bone marrow, one of the 30–40 V genes combine with one of the five J genes and is juxtaposed to a C gene. The recombining process involves deletion of the intervening DNA between the selected genes. **c**, The H chain is encoded by V, D, J and C genes. The assembly of the H chain gene occurs in two stages: one of the D genes joins with a J gene, then one of the V genes joins with that DJ assembly.

were reacted with radioactive probes, one corresponding to the whole L chain, the other to just the C portion. In the tumour, both probes lit up the identical fragment, whereas in the embryonic extract, two different fragments hybridized to the full-length probe, but only one of them to the C-region probe. Both fragments from the embryo sample were different in size from the single hybridizing fragment in the tumour sample. The experiment strongly argued for the V and C genes being some distance from each other in the embryo, but having been rearranged and assembled during development of antibody-forming cells in adults to form a continuous DNA sequence constituting the full L-chain gene.

Generating more diversity

Definitive elucidation of immunoglobulin gene structure depended on molecular cloning and subsequent sequencing of the genes themselves¹³. Here came another surprise, and one that could really not have been anticipated. V-region genes in the germline were found to be significantly shorter than is required to code for the V region of the L chain. It turned out that there is a series of 'minigenes' known as 'joining' (J) genes, which code for about 13 amino acids of the L chain. Thus, the full L chain is actually encoded by V, J and C genes (Fig. 2b). For the H chain, it is still more complicated, as there exists a series of 'diversity' (D) genes that encode up to eight amino acids that lie between the V and J regions. Thus, the H chain is encoded by V, D, J and C genes (Fig. 2c). The assembly of a complete H-chain V region occurs in two separate steps: first, one of the D regions joins with one of the J regions, then one of many V regions joins with that DJ assembly (Fig. 2c). The joining process is followed by deletion of the intervening DNA between the chosen minigenes. This is the first example of a somatic cell possessing a different genome from its fellow cells.

This minigene assembly process has important implications for antibody diversity. In humans, there are two types of L chains, κ and λ , each with its own sets of V and J genes. For the κ light chain, there are 40 functional V genes and 5 functional J genes; for the λ chain, there are 31 and 4, respectively. There is only one kind of variable region for the H chain, encoded by 51 V genes, 25 D genes and 6 J genes. To a first approximation, therefore, there are $(40 \times 5) + (31 \times 4) = 324$ different possible assemblies of L chains, and $51 \times 25 \times 6 = 7,650$ combinations for H chains. Thus, together, there are potentially 2,478,600 different types of germline-encoded antibodies.

But this is a considerable underestimate for two reasons. Recombination junctions can occur at different positions and this junctional diversity increases variability. Furthermore, a few extra nucleotides, called N regions, can be inserted between D–J junctions and V–D junctions in many H chains, and in a smaller percentage of L-chain V–J junctions. These nucleotides are not present in the germline and add to antibody diversity.

Yet further diversity can be generated by DNA mutations in dividing B cells. B cells expressing newly assembled immunoglobulin genes, each with its own unique specificity, constitute the 'primary repertoire'. When an antigen stimulates a chosen B cell to divide (Fig. 1), a proportion of the progeny migrate into the vicinity of antigen-capturing follicular dendritic cells (FDCs) and gradually form a 'germinal centre'. FDCs retain antigen on their surface for long periods and stimulate further rounds of division. Within the germinal centre the B cells display an extraordinarily high rate of somatic mutation in V genes, estimated at 10^{-3} per nucleotide per division¹⁴. As antibody production accumulates, only those B cells with heightened affinity for the antigen gain access to FDC-bound antigen and thus are further stimulated to divide. As a result, B-cell clones secreting higher affinity antibody are selected in an iterative manner.

The 'memory' B cells that emerge from the germinal centre constitute the 'secondary repertoire', which is even more diverse than the primary one. Twenty mutations per chain are not uncommon; nor are thousand-fold increases in affinity. Thus, as it turned out, two early theories of antibody diversification proved to be correct: rearrangement of germline genes gives the naive B-cell

Box 1

Tolerating self

How does the immune system distinguish foreign molecules from the cells of its own body? The failure to form antibodies against self components confounded scientists until Macfarlane Burnet proposed that an antigen introduced during embryonic development would 'trick' the immune system into regarding it as self, so that even in later life it could not react against that antigen²². Peter Medawar's group proved the prediction correct by inducing immunological tolerance to transplantation antigens. Injecting foreign cells from a donor into late-stage embryo mice within the uterus led to a situation where, when these mice had grown up, they could not reject skin grafts from the donor strain²³.

What happens if the antibody generated by a newly created B cell is reactive with a self component? The cell either dies immediately, if the stimulus is one capable of strongly cross-linking the cell-surface immunoglobulin receptors, or else is given a negative signal which impedes its full function²⁴. This negative selection contributes to immunological tolerance.

For a T cell, this is more of a juggling act. Although it must recognize the self-molecule major histocompatibility complex (MHC), it must avoid mounting an abnormal immune response to the host's own cells. For the T-cell repertoire to be fashioned correctly within the thymus, two sorts of selection are required. If the T cell, having assembled its receptor genes, has the capacity to recognize self MHC, it is allowed to mature. This step is known as positive selection. But the T cell must not leave the thymus if it has strong potential to cause autoimmunity. This would be the case if there were too high an affinity of binding with self MHC, or with a peptide from a normal cellular component bound to the MHC groove. Such cells undergo programmed cell death within the thymus, the process of negative selection.

Defects in the immune system's ability to recognize self are the basis of many autoimmune diseases, such as juvenile diabetes and multiple sclerosis.

repertoire, and somatic mutation ensures further diversification during memory B-cell development.

Switching function

There are several different classes of antibodies, all of which have distinct roles that are also produced by rearrangements at the DNA level. There are eight different genes for the C region of the H chain, which specify different antibody functions. Each B cell first links the chosen VDJ assembly to a C gene known as μ , creating an antibody class called IgM. If that cell is propelled into a pathway favouring the production of an antibody prominent in mucus secretions (such as in the gastrointestinal tract), the VDJ section is switched over to a C region encoded by C gene α , and the cell produces IgA. If, on the other hand, the antigen is of parasite origin, or an allergen such as a pollen grain, the cell may be stimulated to produce IgE, in which case the VDJ region associates with the product of the C gene ϵ . All of this occurs without any change in the specificity of the antibody being secreted. Although the detailed molecular mechanisms are still being investigated, the class switching again involves sequential excision of portions of the genetic material. Cytidine deaminase induced by B-cell-specific activation may be significant in both class switching and somatic hypermutation.

Assembly of T-cell receptors

Whereas B cells make antibodies against antigens, the thymus-derived or T cells also respond to foreign agents, specializing in a more localized form of combat. Cytotoxic T cells are capable of killing virus-infected cells or cells displaying cancer-specific antigens. Other

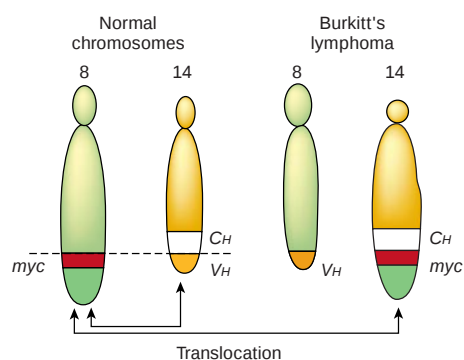


Figure 3 Reciprocal chromosomal translocations in Burkitt's lymphoma, a solid tumour of B lymphocytes. The genes for making the heavy chains of antibodies (C_H) are located on chromosomes 14, whereas those for making the light chains are on chromosomes 2 and 22. These genes are expressed exclusively in B lymphocytes, because only these cells have the necessary transcription factors to switch on their expression. In most (over 90%) of Burkitt's lymphoma cases, a reciprocal translocation moves the proto-oncogene *c-myc* from its normal position on chromosome 8 to a location close to the antibody heavy-chain genes on chromosome 14 (ref. 18). In other cases, *c-myc* is translocated close to the antibody genes on chromosome 2 or 22. In every case, *c-myc* now finds itself in a region of active gene transcription, and it may simply be the overproduction of the *c-myc* product (a transcription factor essential for cell division) that propels the lymphocyte down the pathway towards cancer.

T cells secrete powerful stimulatory and inflammatory molecules, most of which act in a strictly localized context. T cells can also help guide B cells down appropriate pathways of differentiation.

In common with B cells, T cells also have one-receptor specificity and, for simplicity, I shall mention only the $\alpha\beta$ T-cell receptor (TCR), a heterodimer consisting of two subunits, the α - and β -chains, joined by disulphide bonds^{15,16}. The strategy for generating T cells with different receptors is strikingly similar to that used by B cells to produce different antibodies and, in fact, the TCR binding surface looks much like that of an antibody. The β -chain of the TCR is assembled in somatic cells from V, D, J and C genes; the α -chain from V, J and C genes. There are additions of N-region nucleotides between V and D, as well as between D and J on the β -chain; and between V and J on the α -chain. A similar rearrangement also takes place for the $\gamma\delta$ TCR.

But something peculiar about T-cell recognition was noted by Rolf Zinkernagel and Peter Doherty¹⁷, who demonstrated that cytotoxic T cells could recognize viral antigens only if a specific 'self' molecule were also present on the target cell (see Box 1). The key part of the T-cell recognition puzzle fell into place when it was discovered that the TCR recognized short antigenic peptides bound to the groove of a self molecule known as the major histocompatibility complex (MHC), as well as surrounding portions of the MHC molecule itself. Cells have special mechanisms for fragmenting proteins into peptides of 8–24 amino acids in length, attaching these to MHC molecules and transporting the entire complex to the cell surface. TCRs then 'see' these short linear portions of antigens, be these of viral, bacterial or parasitic origin, or even portions of normal intracellular components. Such a system can help to control infections where the pathogen goes 'underground' inside a cell, and can also eliminate cells with mutated self antigens, such as cancer cells.

Lymphocytes and cancer

Lymphocytes have been a favourite tool in cancer research. A notable example of DNA science applied in this way relates to the B-cell tumour of humans known as Burkitt's lymphoma. Occasionally DNA strands break and are incorrectly repaired. Thus, a piece of a chromosome becomes attached to the broken end of another one, and vice versa, in a process known as reciprocal translocation (Fig. 3).

In the case of Burkitt's lymphoma, a tumour-promoting gene or oncogene called *myc* is translocated from its normal position on chromosome 8 right into the middle of the IgH chain locus on chromosome 14 (ref. 18). In this highly active transcriptional environment, *myc* expression is switched on, and eventually cancer develops.

It has been possible to create lymphoma-prone transgenic mice, which express *myc* in aberrantly high amounts. Because cancer is typically a multistage process, if further oncogenes are expressed simultaneously in transgenic mice, the onset of cancer can be dramatically accelerated. One such example involves the gene *bcl-2*. When this gene is expressed, it stops cells from undergoing natural programmed death (apoptosis)¹⁹. Mice expressing *myc* and *bcl-2* showed very rapid development of tumours. An enormous amount of literature has accumulated related to the expanding family of *bcl-2*-related genes and their roles in the regulation of programmed cell death. Models derived from lymphocytes and their malignancies have led to insights with implications well beyond immunology.

DNA vaccines

DNA research has been of immense value to vaccine research. Through gene cloning and expression, candidate antigens can be identified and tested. In an era of rapid nucleotide sequencing, the whole genome of a pathogen can be determined, and computer programs can search for sequences likely to encode outer membrane proteins, which can be assessed as candidate vaccine molecules (see, for example, the series of papers published recently in *Nature* (419, 489–542, 2002) on the genomics of the malaria parasite).

Amazingly, DNA itself can serve as a vaccine. DNA vaccines work on the principle that the gene sequence for one or more candidate antigens is introduced into an animal or person via a delivery vehicle known as a vector, together with a strong promoter that can switch on its expression in mammalian cells. Cells that take up the injected DNA transcribe and translate the gene and release the relevant antigen protein, which the body can in turn manufacture antibodies against. Thus, the body itself becomes a vaccine factory²⁰. Unfortunately, so far this approach has worked better in mice than in humans, but many avenues are being pursued to improve this situation.

To strengthen the immune response to a vaccine, it may be necessary to use an adjuvant substance. Here, DNA may also be of potential use. Scavenger cells, which capture antigens, have evolutionarily conserved receptors, known as Toll-like receptors (TLRs), which recognize antigens common to many pathogens. One such receptor is TLR-9, which recognizes unmethylated CpG motifs commonly found in bacterial but not mammalian DNA. Accordingly, unmethylated CpG-rich DNA sequences represent a promising new category of adjuvant²¹.

Future directions

The solutions to the puzzle of antibody diversity and mystery of T-cell recognition of antigenic peptides are among the brightest chapters of biology in the last quarter of the twentieth century. The future of immunology will be all about how the system is regulated and how it makes decisions: whether to respond or not; whether to direct efforts towards antibody formation or cell-mediated immunity; and, if the latter, whether more towards cytokine-secreting T cells or cytotoxic T cells.

As in the past, the future will be about information and thus about DNA science. All the complex signalling pathways, the feedback loops, and the intricate rules governing cell division on the one hand or programmed cell death on the other, will be progressively revealed. As this happens, the possibilities for applied research and development will be immense. In particular, new therapeutic targets will be identified. The 'miracle' drug for chronic myelogenous leukaemia, Glivec, was made possible after the characterization of the extraordinary cancerous potential of the chimaeric oncogene *bcr-abl*. This will surely be only the first of a plethora of more intelligently designed

anti-cancer drugs. Potent cytokines and monoclonal antibodies directed against cell surface-associated structures are already prominent within a radically revised pharmaceutical armamentarium in areas including cancer, autoimmunity, allergy and transplantation. DNA research is therefore crucial to a new generation of immunologists, from those striving towards the development of novel vaccines to those seeking to understand and control autoimmune diseases, allergy and transplant tolerance. □

doi:10.1038/nature01409

1. Ehrlich, P. The Croonian Lecture. On immunity with special reference to cell life. *Proc. R. Soc. Lond. B* **66**, 424–448 (1900).
2. Landsteiner, K. *The Specificity of Serological Reactions* (Harvard Univ. Press, Boston, 1945).
3. Breinl, F. & Haurowitz, F. Untersuchungen des Präzipitates aus Hämoglobin und anti-Hämoglobin-Serum und Bemerkungen über die Natur der Antikörper. *Hoppe-Seyler's Z. Physiol. Chem.* **192**, 45–57 (1930).
4. Jerne, N. K. The natural selection theory of antibody formation. *Proc. Natl Acad. Sci. USA* **41**, 849–857 (1955).
5. Talmage, D. W. Allergy and immunology. *Annu. Rev. Med.* **8**, 239–256 (1957).
6. Burnet, F. M. A modification of Jerne's theory of antibody production using the concept of clonal selection. *Aust. J. Sci.* **20**, 67–69 (1957).
7. Nossal, G. J. V. & Lederberg, J. Antibody production by single cells. *Nature* **181**, 1419–1420 (1958).
8. Edelman, G. M. & Gall, W. E. The antibody problem. *Annu. Rev. Biochem.* **38**, 415–466 (1969).
9. Hilschmann, N. & Craig, L. C. Amino acid sequence studies with Bence-Jones proteins. *Proc. Natl Acad. Sci. USA* **53**, 1403–1409 (1965).
10. Dreyer, W. J. & Bennett, J. C. The molecular basis of antibody formation: a paradox. *Proc. Natl Acad. Sci. USA* **54**, 864–869 (1965).
11. Wu, T. T. & Kabat, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211–250 (1970).
12. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl Acad. Sci. USA* **73**, 3628–3632 (1976).
13. Bernard, O., Hozumi, N. & Tonegawa, S. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* **15**, 1133–1144 (1978).
14. Kocks, C. & Rajewsky, K. Stable expression and somatic hypermutation of antibody V regions in B-cell developmental pathways. *Annu. Rev. Immunol.* **7**, 537–559 (1989).
15. Hedrick, S. M., Cohen, D. I., Nielsen, E. A. & Davis, M. M. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* **308**, 149–153 (1984).
16. Yanagi, Y. *et al.* A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature* **308**, 145–149 (1984).
17. Zinkernagel, R. M. & Doherty, P. C. Restriction of *in vitro* T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* **248**, 701–702 (1974).
18. Bernard, O., Cory, S., Gerondakis, S., Webb, E. & Adams, J. M. Sequence of the murine and human cellular myc oncogenes and two modes of myc transcription resulting from chromosome translocation in B lymphoid tumours. *EMBO J.* **2**, 2375–2383 (1983).
19. Vaux, D. L., Cory, S. & Adams, J. M. *Bcl-2* gene promotes haemopoietic cell survival and cooperates with *c-myc* to immortalize pre-B cells. *Nature* **335**, 440–442 (1988).
20. Donnelly, J. J., Ulmer, J. B., Shiver, J. W. & Liu, M. A. DNA vaccines. *Annu. Rev. Immunol.* **15**, 617–648 (1997).
21. Krieg, A. M. CpG motifs in bacterial DNA and their immune effects. *Annu. Rev. Immunol.* **20**, 709–760 (2002).
22. Burnet, F. M. & Fenner, F. J. *The Production of Antibodies* (Macmillan, Melbourne, 1949).
23. Billingham, R. F., Brent, L. & Medawar, P. B. Actively acquired tolerance of foreign cells. *Nature* **172**, 603–606 (1953).
24. Nossal, G. J. V. & Pike, B. L. Clonal anergy: persistence in tolerant mice of antigen-binding B lymphocytes incapable of responding to antigen or mitogen. *Proc. Natl Acad. Sci. USA* **77**, 1602–1606 (1980).

The digital code of DNA

Leroy Hood* & David Galas†

*Institute for Systems Biology, 4225 Roosevelt Way NE, Seattle, Washington 98105, USA (e-mail: lhood@systemsbiology.org)

†Kek Graduate Institute of Applied Sciences, 535 Watson Drive, Claremont, California 91711, USA (e-mail: david_galas@kgi.edu)

The discovery of the structure of DNA transformed biology profoundly, catalysing the sequencing of the human genome and engendering a new view of biology as an information science. Two features of DNA structure account for much of its remarkable impact on science: its digital nature and its complementarity, whereby one strand of the helix binds perfectly with its partner. DNA has two types of digital information — the genes that encode proteins, which are the molecular machines of life, and the gene regulatory networks that specify the behaviour of the genes.

"Any living cell carries with it the experiences of a billion years of experimentation by its ancestors." Max Delbruck, 1949.

The discovery of the double helix in 1953 immediately raised questions about how biological information is encoded in DNA¹. A remarkable feature of the structure is that DNA can accommodate almost any sequence of base pairs — any combination of the bases adenine (A), cytosine (C), guanine (G) and thymine (T) — and, hence any digital message or information. During the following decade it was discovered that each gene encodes a complementary RNA transcript, called messenger RNA (mRNA)², made up of A, C, G and uracil (U), instead of T. The four bases of the DNA and RNA alphabets are related to the 20 amino acids of the protein alphabet by a triplet code — each three letters (or 'codons') in a gene encodes one amino acid³. For example, AGT encodes the amino acid serine. The dictionary of DNA letters that make up the amino acids is called the genetic code⁴. There are 64 different triplets or codons, 61 of which encode an amino acid (different triplets can encode the same amino acid), and three of which are used for 'punctuation' in that they signal the termination of the growing protein chain.

The molecular complementarity of the double helix — whereby each base on one strand of DNA pairs with its complementary base on the partner strand (A with T, and C with G) — has profound implications for biology. As implied by James Watson and Francis Crick in their landmark paper¹, base pairing suggests a template-copying mechanism that accounts for the fidelity in copying of genetic material during DNA replication (see article in this issue by Alberts, page 431). It also underpins the synthesis of mRNA from the DNA template, as well as processes of repairing damaged DNA (discussed by Friedberg, page 436).

Tools to modify DNA

The enzymes that function in cells to copy, cut and join DNA molecules were also exploited as key tools for revolutionary new techniques in molecular biology, including the cloning of genes and expression of their proteins, and mapping the location of genes on chromosomes. The ability to recreate the process of DNA replication artificially in the laboratory led to the development of two techniques



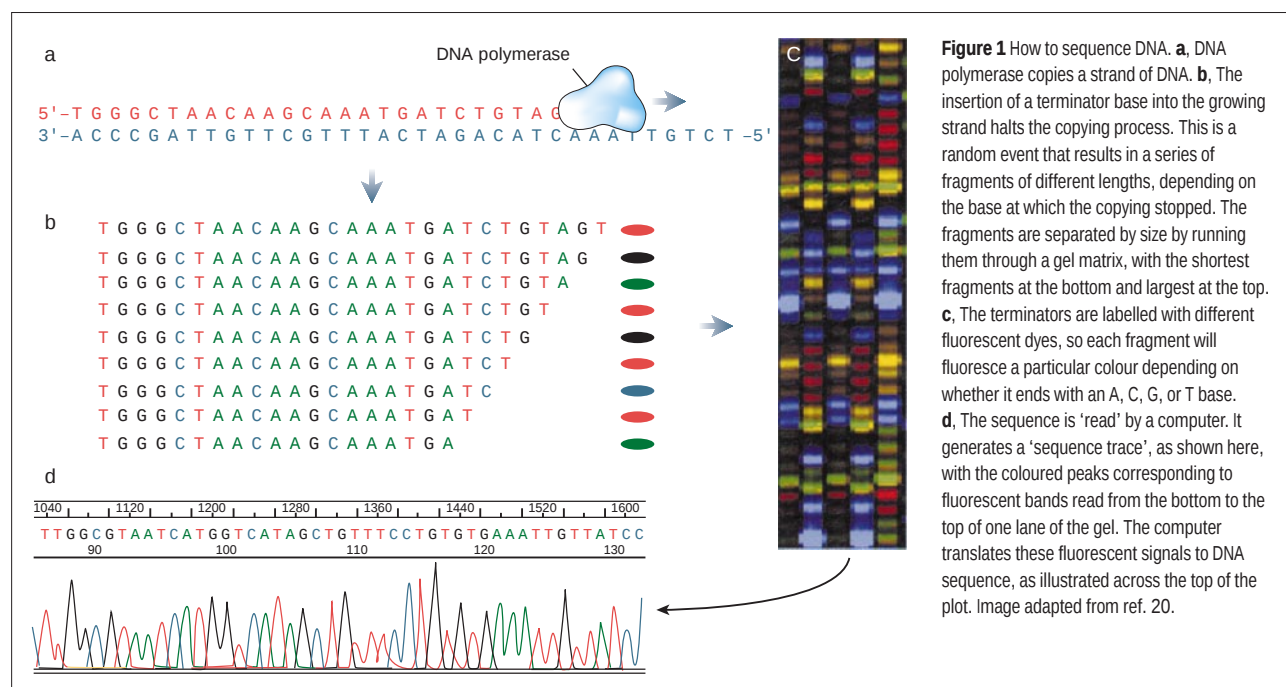


Figure 1 How to sequence DNA. **a**, DNA polymerase copies a strand of DNA. **b**, The insertion of a terminator base into the growing strand halts the copying process. This is a random event that results in a series of fragments of different lengths, depending on the base at which the copying stopped. The fragments are separated by size by running them through a gel matrix, with the shortest fragments at the bottom and largest at the top. **c**, The terminators are labelled with different fluorescent dyes, so each fragment will fluoresce a particular colour depending on whether it ends with an A, C, G, or T base. **d**, The sequence is 'read' by a computer. It generates a 'sequence trace', as shown here, with the coloured peaks corresponding to fluorescent bands read from the bottom to the top of one lane of the gel. The computer translates these fluorescent signals to DNA sequence, as illustrated across the top of the plot. Image adapted from ref. 20.

that transformed biology: a manual DNA sequencing method in 1975 and, in 1985, the discovery of the polymerase chain reaction (PCR), whereby DNA sequences could be amplified a millionfold or more⁵.

Although sequencing and PCR transformed the science of biology, they also had wide applications for medicine and forensics. The detection of variation in DNA sequence from one individual to the next — so-called 'polymorphisms' — forms the basis of DNA 'finger-printing' of individuals. Forensics uses these fingerprints to deal with paternity disputes, as well as criminal cases such as rape. The finding that many specific DNA polymorphisms are associated with disease or disease susceptibility has brought DNA diagnostics to medicine and opened the pathway to truly predictive medicine, where the risks of disease can be identified in advance of symptoms (see article in this issue by Bell, page 414).

Automated DNA sequencing

The first efforts to sequence DNA, pioneered by Walter Gilbert⁶ and Fred Sanger⁷ in the 1970s, decoded stretches of DNA a few hundred bases long. When the first complete genome was sequenced over a period of about one year in 1977–78 — that of a viral genome of about 5,000 bases⁸ — it became clear that DNA sequence data could provide unique insights into the structure and function of genes, as well as genome organization. It was this potential to generate vast amounts of information about an organism from its genetic code that inspired efforts towards the automation of DNA sequencing (Fig. 1).

The combination of technical wizardry and intensive automation in the decade that followed launched the 'genomic era'. A series of new instruments enabled novel approaches to biological analysis^{9–11}. The first sequencing machine — invented by Leroy Hood, Lloyd Smith and Mike Hunkapiller in 1986 (ref. 12) — was automated in data acquisition, but still required substantial manual attention and the sequencing rate was low, roughly 250 bases per day. Over the next ten years, the development of automated DNA sequencing accelerated, rapidly passing through three distinct stages: the prototype sequencing machine (1986); a robust instrument that could be used routinely in a standard laboratory (1989); and finally, a machine that formed part of an integrated factory-like production line where DNA sample preparation and sequencing were all fully automated (1998). The

advances in sequencing capacity have been striking — the latest sequencing machines are able to decode approximately 1.5 million bases over 24 hours — 6,000 times the throughput of the prototype.

The goals of high-throughput biological instrumentation are to increase throughput, enhance the quality of the data, and greatly reduce the cost of per unit information acquired. To reach these goals in the future, the miniaturization, automation, parallelization and integration of successive procedures will propel DNA sequencing technology into the realm of microfluidics and microelectronics, and eventually into the area of nanotechnology. With single-DNA-molecule sequencing, we foresee a time when the entire genome of an individual could be sequenced in a single day at a cost of less than \$US10,000 (compared with the US\$50 million or more it would cost today). This will readily enable the decoding of the genomic sequence of virtually any organism on the planet and provide unparalleled access to the foundations of biology and the study of human genetic variability.

The Human Genome Project

The breathtaking speed at which automated DNA sequencing developed was largely stimulated by the throughput demands of the Human Genome Project (HGP), which officially started in 1990 following discussions and studies on feasibility and technology that began in earnest in 1985. The objectives of the HGP were to generate a finished sequence in 15 years¹³, but a draft of the human genome sequence was available in 2001. Two versions of the draft were generated and published in 2001, one by the publicly funded International Human Genome Sequencing Consortium¹⁴, and another by the biotechnology company Celera¹⁵ (Box 1). In the process of developing the tools and methodology to be able to sequence and assemble the 3 billion bases of the human genome, a range of plant, animal and microbial genomes was sequenced and many more are currently being decoded. As genome sequences become available, different areas of biology are being transformed — for example, the discipline of microbiology has changed significantly with the completion of more than 100 bacterial genome sequences over the past decade.

The HGP profoundly influenced biology in two respects. First, it illustrated the concept of 'discovery science' — the idea that all the elements of the system (that is, the complete genome sequence and

the entire RNA and protein output encoded by the genome) can be defined, archived in a database, and made available to facilitate hypothesis-driven science and global analyses. Second, to succeed, the HGP pushed the development of efficient large-scale DNA sequencing and, simultaneously, drove the creation of high-throughput tools (for example, DNA arrays and mass spectrometry) for the analysis of other types of related biological information, such as mRNAs, proteins and molecular interactions.

The digital nature of biological information

The value of having an entire genome sequence is that one can initiate the study of a biological system with a precisely definable digital core of information for that organism — a fully delineated genetic source code. The challenge, then, is in deciphering what information is encoded within the digital code. The genome encodes two main types of digital information — the genes that encode the protein and RNA molecular machines of life, and the regulatory networks that specify how these genes are expressed in time, space and amplitude.

It is the evolution of the regulatory networks and not the genes themselves that play the critical role in making organisms different from one another. The digital information in genomes operates across three diverse time spans: evolution (tens to millions of years), development (hours to tens of years), and physiology (milliseconds to weeks). Development is the elaboration of an organism from a single cell (the fertilized egg) to an adult (for humans this is 10^{14} cells of thousands of different types). Physiology is the triggering of specific functional programmes (for example, the immune response) by environmental cues. Regulatory networks are crucial in each of these aspects of biology.

Regulatory networks are composed of two main types of components: transcription factors and the DNA sites to which they bind in the control regions of genes, such as promoters, enhancers and silencers. The control regions of individual genes serve as information processors to integrate the information inherent in the concentrations of different transcription factors into signals that mediate gene expression. The collection of the transcription factors and their cognate DNA-binding sites in the control regions of genes that carry out a particular developmental or physiological function constitute these regulatory networks (Fig. 2).

Because most 'higher' organisms or eukaryotes (organisms that contain their DNA in a cellular compartment called the nucleus), such as yeast, flies and humans, have predominantly the same families of genes, it is the reorganization of DNA-binding sites in the control regions of genes that mediate the changes in the developmental programmes that distinguish one species from another. Thus, the regulatory networks are uniquely specified by their DNA-binding sites and, accordingly, are basically digital in nature.

One thing that is striking about digital regulatory networks is that they can change significantly in short periods of evolutionary time. This is reflected, for example, in the huge diversity of the body plans, controlled by gene regulatory networks, that emerged over perhaps 10–30 million years during the Cambrian explosion of metazoan organisms (about 550 million years ago). Likewise, remarkable changes occurred to the regulatory networks driving the development of the human brain during its divergence from its common ancestor with chimpanzees about 6 million years ago.

Biology has evolved several different types of informational hierarchies. First, a regulatory hierarchy is a gene network that defines the relationships of a set of transcription factors, their DNA-binding sites and the downstream peripheral genes that collectively control a particular aspect of development. A model of development in the sea urchin represents a striking example¹⁶ (Fig. 2). Second, an evolutionary hierarchy defines an order set of relationships, arising from DNA duplication. For example, a single gene may be duplicated to generate a multi-gene family, and a multi-gene family may be duplicated to create a supergene family. Third, molecular machines may be assembled into structural hierarchies by an ordered assembly process. One

Box 1

Sequencing the human genome

The first complete drafts of the human genome sequence were published in 2001 by the International Human Genome Sequencing Consortium (IHGSC), a publicly funded effort, and Celera, a biotechnology company, using different approaches. Both efforts used a random or shotgun approach where the original DNA to be sequenced was randomly broken into overlapping fragments that were then cloned, and 500 base pairs (bp) were 'read' from one or both ends of the clones.

For the draft genome sequences, each base was read six to ten times to optimize the accuracy of the sequence. The stretches of DNA sequence were read by a computer and assembled into a complete sequence. The IHGSC effort randomly cleaved DNA into ~200,000-bp fragments and generated a map of these fragments across the 24 different human chromosomes; it then used the shotgun approach to sequence the pre-ordered fragments clone by clone. In contrast, Celera randomly fragmented the entire genome into three sizes of fragments (approximately 2,000, 10,000 and 200,000 bp), sequenced both ends of the clones and then used the end sequences to assemble the entire genome sequence, without the aid of a map.

Celera's 1998 announcement that it would sequence the human genome within three years was greeted with considerable scepticism, but it succeeded in producing a draft sequence and considerably accelerating the public effort. The efforts of both groups benefited science by producing draft genome sequences considerably earlier than expected.

Although minor differences were noted between the two drafts, the overall conclusions concerning gene numbers, repeated sequences and chromosomal organization were remarkably similar. For example, both groups identified 30,000–35,000 genes, far fewer than the 100,000 expected from an earlier (admittedly 'back of the envelope') calculation.

example of this is the basic transcription apparatus that involves the step-by-step recruitment of factors and enzymes that will ultimately drive the specific expression of a given gene. A second example is provided by the ribosome, the complex that translates RNA into protein, which is assembled from more than 50 different proteins and a few RNA molecules. Finally, an informational hierarchy depicts the flow of information from a gene to environment: gene → RNA → protein → protein interactions → protein complexes → networks of protein complexes in a cell → tissues or organs → individual organisms → populations → ecosystems. At each successively higher level in the informational hierarchy, information can be added or altered for any given element (for example, by alternative RNA splicing or protein modification).

Systems approaches to biology

Humans start life as a single cell — the fertilized egg — and develop into an adult with trillions of cells and thousands of cell types. This process uses two types of biological information: the digital information of the genome, and environmental information, such as metabolite concentrations, secreted or cell-surface signals from other cells or chemical gradients. Environmental information is of two distinct types: deterministic information where the consequences of the signals are essentially predetermined, and stochastic information where chance dictates the outcome.

Random, or stochastic, signals can generate significant noise in biological systems, but it is only in special cases that noise is converted into signals. For example, stochastic events govern many of the genetic mechanisms responsible for generating antibody diversity. In the immune response, those B cells that produce antibodies that bind

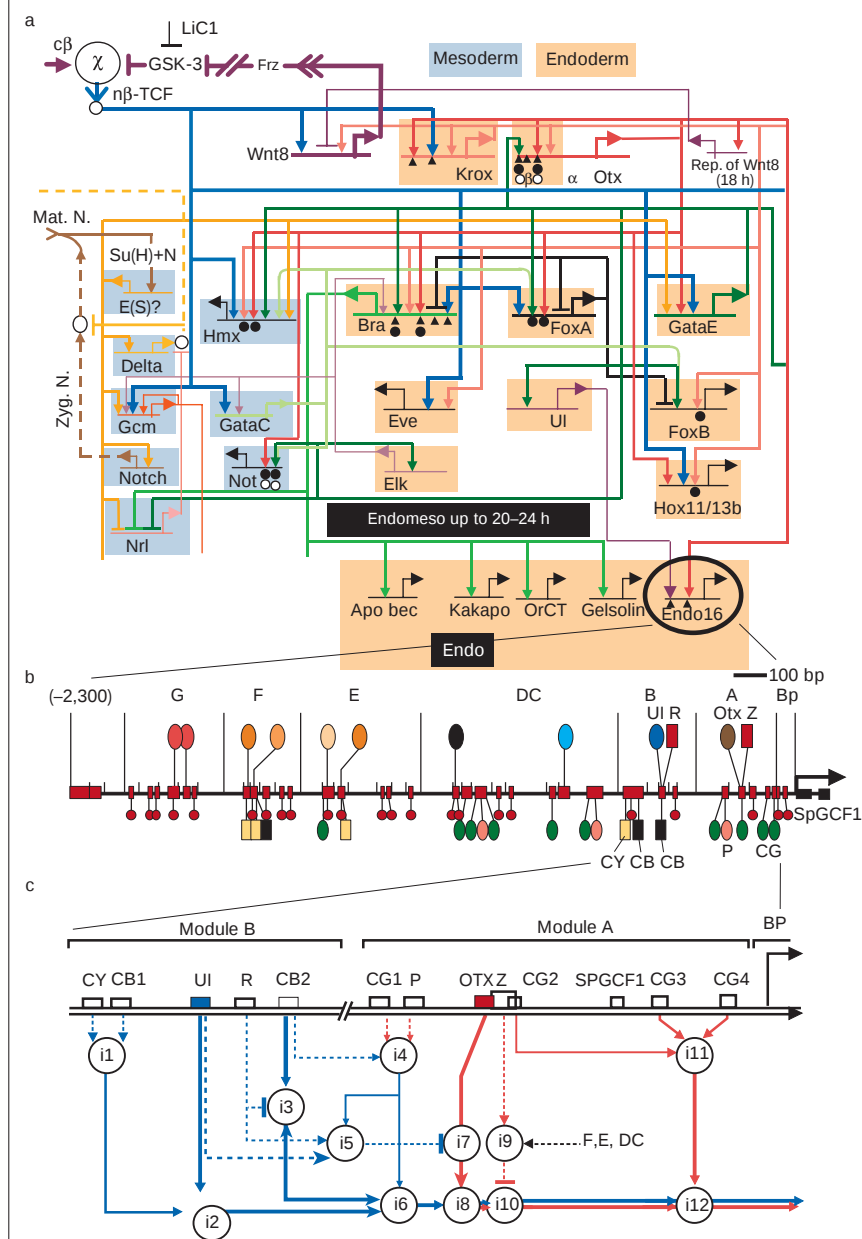


Figure 2 A gene regulatory network involved in sea urchin development¹⁶. **a**, Part of the network of transcription factors and their interactions with the control regions of other transcription factors. Genes are indicated by horizontal lines; arrowheads indicate activation; '⊥' symbols indicate gene repression. **b**, An enlargement of the promoter region of a gene, called *endo 16*, that helps modulate the development of the endoderm. It contains 34 binding sites (rectangles) for 13 different transcription factors and cofactors (illustrated as rectangles or lollipops, respectively). Six modules (A–G) of transcription factors and binding sites carry out discrete functions to developmentally regulate *endo 16*. **c**, Diagram depicting the logical structures of the A and B control circuits during sea urchin development.

tightly to the antigen (that is, those having high affinities) undergo an expansion in number that is proportional to the strength of the antibody affinity (see article in this issue by Nossal, page 440). Hence, the signal (high affinity) is distinguished from the noise (low affinity). Moreover, high levels of mutation in the B cells causes specific diversification of antibody genes in the presence of antigen and permits the affinity to increase even more. The cells carrying the higher-affinity antibody genes are then preferentially selected for survival and proliferation.

The key question is what and how much signal emerges from the noise. Analysis of stochastic events and the differentiation between signal and noise will be a future challenge for contemporary biology. The immune response has been studied for more than 100 years, yet we still have only a partial understanding of its systems properties, such as the immune response and tolerance (the unresponsiveness to one's own cells). This is because until recently immunologists have been able to study this complex system only one gene or one protein at a time.

The systems approach permits the study of all elements in a system in response to genetic (digital) or environmental perturbations. Global quantitative analyses of biological information from different levels each provide new insights into the operation of the system; hence, information at as many levels as possible must be captured, integrated, and ultimately, modelled mathematically. The model should explain the properties of the system and establish a framework that allows us to redesign the system in a rational way to generate new emergent properties.

Several systems have been explored successfully. The utilization of the sugar galactose in yeast has been analysed using genetic perturbations (inactivation of genes) and four levels of information were gathered — RNA and protein concentrations as well as protein–protein and protein–DNA interactions¹⁷. Using an iterative and integrative systems approach, new insights into the regulation of galactose use were gained. Moreover, the relationships of the galactose regulatory network to other modules in the yeast cell were also delineated. Likewise, systems approaches to early embryonic

development in the sea urchin have delineated a regulatory network that has significant predictive power¹⁶ (Fig. 2). Finally, systems approaches to metabolism in an archaeal halobacterium (an organism thriving in up to five-molar salt solutions, such as the Dead Sea) have revealed new insights into the inter-relationships among several modules controlling energy production in the cell¹⁸.

The study of cellular and organismal biology using the systems approach is at its very beginning. It will require integrated teams of scientists from across disciplines — biologists, chemists, computer scientists, engineers, mathematicians and physicists. New methods for acquiring and analysing high-throughput biological data are needed. A powerful computational infrastructure must be leveraged to generate more effective approaches to the capture, storage, analysis, integration, graphical display and mathematic formulation of biological complexity. New technologies must be integrated with each other. Finally, hypothesis-driven and discovery science must be integrated. In short, both new science and technology must emerge for the systems biology approach to realize its promise. A cultural shift in the biological sciences is needed, and the education and training of the next generation of biologists will require significant reform.

Gordon Moore, the founder of Intel, predicted that the number of transistors that could be placed on a computer chip would double every 18 months. It has for more than 30 years. This exponential growth has been a driver for the explosive growth of information technology. Likewise, the amount of DNA sequence information available to the scientific community is following a similar, perhaps even steeper, exponential increase. The critical issue is how sequence information can be converted into knowledge of the organism and how biology will change as a result. We believe that a systems approach to biology is the key. It is clear, however, that this approach poses significant challenges, both scientific and cultural¹⁹. The discovery of DNA structure started us on this journey, the end of which will be the grand unification of the biological sciences in the emerging, information-based view of biology. □

doi:10.1038/nature01410

1. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 576–581 (1961).
3. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
4. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polynucleotides. *Proc. Natl Acad. Sci. USA* **47**, 1588–1602 (1961).
5. Saiki, R. K. *et al.* Enzymatic amplification of β -globin sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
6. Maxam, A. M. & Gilbert, W. A new method of sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977).
7. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 444–448 (1975).
8. Sanger, F. *et al.* Nucleotide sequence of bacteriophage ϕ X174. *Nature* **265**, 678–695 (1977).
9. Hunkapiller, M. W. & Hood, L. New protein sequenator with increased sensitivity. *Science* **207**, 523–525 (1980).
10. Horvath, S. J., Firca, J. R., Hunkapiller, T., Hunkapiller, M. W. & Hood, L. An automated DNA synthesizer employing deoxynucleoside 3' phosphoramidites. *Methods Enzymol.* **154**, 314–326 (1987).
11. Kent, S. B. H., Hood, L. E., Beilan, H., Meister, S. & Geiser, T. High yield chemical synthesis of biologically active peptides on an automated peptide synthesizer of novel design. *Peptides* **5**, 185–188 (1984).
12. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
13. Collins, F. & Galas, D. J. A new five-year plan for the US Human Genome Project. *Science* **262**, 43–46 (1993).
14. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
15. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
16. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
17. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–933 (2001).
18. Baliga, N. S. *et al.* Coordinate regulators of energy transduction modules in the *Halobacterium* sp. analyzed by a global systems approach. *Proc. Natl Acad. Sci. USA* **99**, 14913–14918 (2002).
19. Aderem, A. & Hood, L. Immunology in the post-genomic era. *Nature Immunol.* **2**, 1–3 (2001).
20. Dennis, C. & Gallagher, R. (eds) *The Human Genome* (Palgrave, Basingstoke, 2001).

Controlling the double helix

Gary Felsenfeld* & Mark Groudine†

*Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, Room 212, Bethesda, Maryland 20892-0540, USA (e-mail: gary.felsenfeld@nih.gov)

†Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, and Department of Radiation Oncology, University of Washington School of Medicine, Seattle, Washington 98195, USA (e-mail: markg@fhccr.org)

Chromatin is the complex of DNA and proteins in which the genetic material is packaged inside the cells of organisms with nuclei. Chromatin structure is dynamic and exerts profound control over gene expression and other fundamental cellular processes. Changes in its structure can be inherited by the next generation, independent of the DNA sequence itself.

Genes were first shown to be made of DNA only nine years before the structure of DNA was discovered (ref. 1; and see article in this issue by McCarty, page 406). Although revolutionary, the idea that genetic information was protein-free ultimately proved too simple. DNA in organisms with nuclei is in fact coated with at least an equal mass of protein, forming a complex called chromatin, which controls gene activity and the inheritance of traits.

'Higher' organisms, such as yeast and humans, are eukaryotes; that is, they package their DNA inside cells in a separate compartment called the nucleus. In dividing cells, the chromatin complex of DNA and protein can be seen as individual compact chromosomes; in non-dividing cells, chromatin appears to be distributed throughout the nucleus and organized into 'condensed' regions (heterochromatin) and more open 'euchromatin' (see article in this issue by Ball, page 421). In contrast, prokaryotes, such as bacteria, lack nuclei.

The evolution of chromatin

The principal protein components of chromatin are proteins called histones (Fig. 1). Core histones are among the most highly conserved eukaryotic proteins known, suggesting that the fundamental structure of chromatin evolved in a common ancestor of eukaryotes. Moreover, histone equivalents and a simplified chromatin structure have also been found in single-cell organisms from the kingdom Archaeobacteria^{2,3}.

Because there is more DNA in a eukaryote than in a prokaryote, it was naturally first assumed that the purpose of histones was to compress the DNA to fit within the nucleus. But subsequent research has dramatically revised the view that histones emerged as an afterthought, forced on eukaryotic DNA as a consequence of large genome size and the constraints of the nucleus.

It was known that different genes are active in different tissues, and the distinction of heterochromatin and euchromatin suggested that differences in chromatin structure were associated with differences in gene expression. This led to the early supposition that the histones were also repressor proteins designed to shut off unwanted expression. The available evidence, although rudimentary, does indeed suggest that archaeal histones are not merely packaging factors, but function to regulate gene expression^{2–5}. They

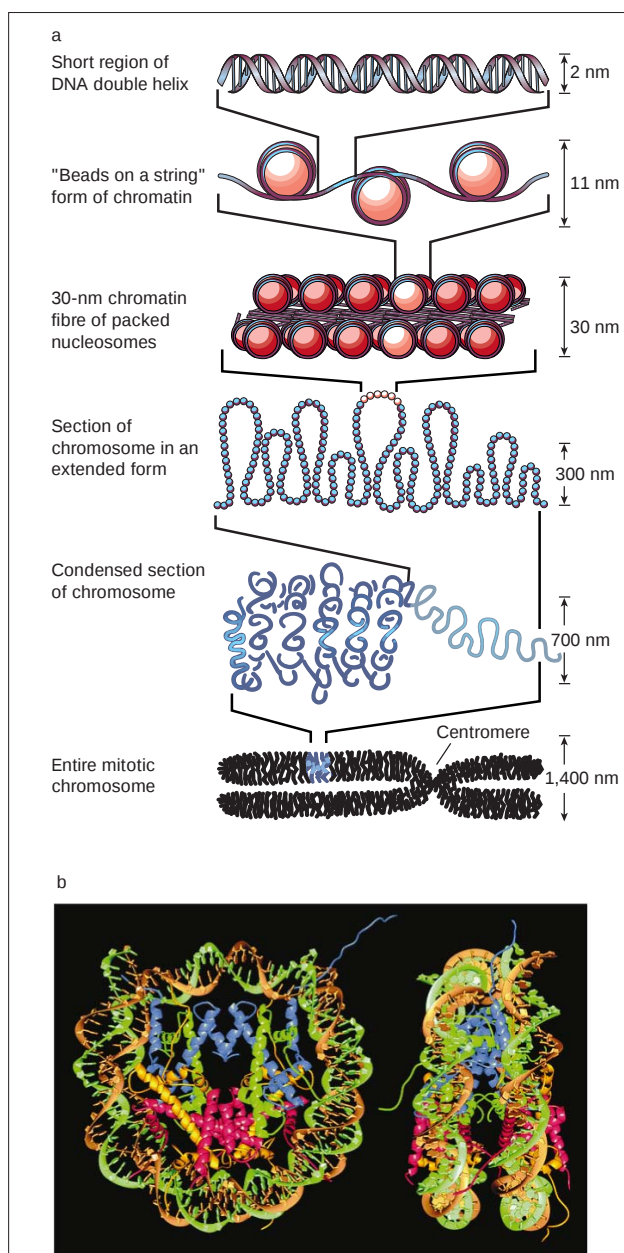


Figure 1 Packaging DNA. **a**, The organization of DNA within the chromatin structure. The lowest level of organization is the nucleosome, in which two superhelical turns of DNA (a total of 165 base pairs) are wound around the outside of a histone octamer. Nucleosomes are connected to one another by short stretches of linker DNA. At the next level of organization the string of nucleosomes is folded into a fibre about 30 nm in diameter, and these fibres are then further folded into higher-order structures. At levels of structure beyond the nucleosome the details of folding are still uncertain. (Redrawn from ref. 41, with permission). **b**, The structure of the nucleosome core particle was uncovered by X-ray diffraction, to a resolution of 2.8 Å (ref. 42). It shows the DNA double helix wound around the central histone octamer. Hydrogen bonds and electrostatic interactions with the histones hold the DNA in place.

may facilitate gene activation, by promoting specific structural interactions between distal sequences, or repression, by occluding binding sites for transcriptional activators.

We suggest that the function of archaeal histones reflects their ancestral function, and therefore that chromatin evolved originally as an important mechanism for regulating gene expression. Its use in

packaging DNA was an ancillary benefit that was recruited for the more complex nucleosome structure that subsequently evolved in the ancestors of modern eukaryotes, which had expanded genome sizes. Although their compactness might seem to suggest inertness, chromatin structures are in fact a centre for a range of biochemical activities that are vital to the control of gene expression, as well as DNA replication and repair.

Packaging DNA into chromatin

The fundamental subunit of chromatin is the nucleosome, which consists of approximately 165 base pairs (bp) of DNA wrapped in two superhelical turns around an octamer of core histones (two each of histones H2A, H2B, H3 and H4). This results in a five- to tenfold compaction of DNA⁶. The DNA wound around the surface of the histone octamer (Fig. 1) is partially accessible to regulatory proteins, but could become more available if the nucleosome could be moved out of the way, or if the DNA partly unwound from the octamer. The histone 'tails' (the amino-terminal ends of the histone protein chains) are also accessible, and enzymes can chemically modify these tails to promote nucleosome movement and unwinding, with profound local effects on the chromatin complex.

Each nucleosome is connected to its neighbours by a short segment of linker DNA (~10–80 bp in length) and this polynucleosome string is folded into a compact fibre with a diameter of ~30 nm, producing a net compaction of roughly 50-fold. The 30-nm fibre is stabilized by the binding of a fifth histone, H1, to each nucleosome and to its adjacent linker. There is still considerable debate about the finer points of nucleosome packing within the chromatin fibre, and even less is known about the way in which these fibres are further packed within the nucleus to form the highest-order structures.

Chromatin regulates gene expression

Regulatory signals entering the nucleus encounter chromatin, not DNA, and the rate-limiting biochemical response that leads to activation of gene expression in most cases involves alterations in chromatin structure. How are such alterations achieved?

The most compact form of chromatin is inaccessible and therefore provides a poor template for biochemical reactions such as transcription, in which the DNA duplex must serve as a template for RNA polymerase. Nucleosomes associated with active genes were shown to be more accessible to enzymes that attack DNA than those associated with inactive genes⁷, which is consistent with the idea that activation of gene expression should involve selective disruption of the folded structure.

Clues as to how chromatin is unpacked came from the discovery that components of chromatin are subject to a wide range of modifications that are correlated with gene activity. Such modifications probably occur at every level of organization, but most attention has focused on the nucleosome itself. There are three general ways in which chromatin structure can be altered. First, nucleosome remodelling can be induced by complexes designed specifically for the task⁸; this typically requires that energy be expended by hydrolysis of ATP. Second, covalent modification of histones can occur within the nucleosome⁹. Third, histone variants may replace one or more of the core histones^{10–12}.

Some modifications affect nucleosome structure or lability directly, whereas others introduce chemical groups that are recognized by additional regulatory or structural proteins. Still others may be involved in disruption of higher-order structure. In some cases, the packaging of particular genes in chromatin is required for their expression¹³. Thus, chromatin can be involved in both activation and repression of gene expression.

Chromatin remodelling

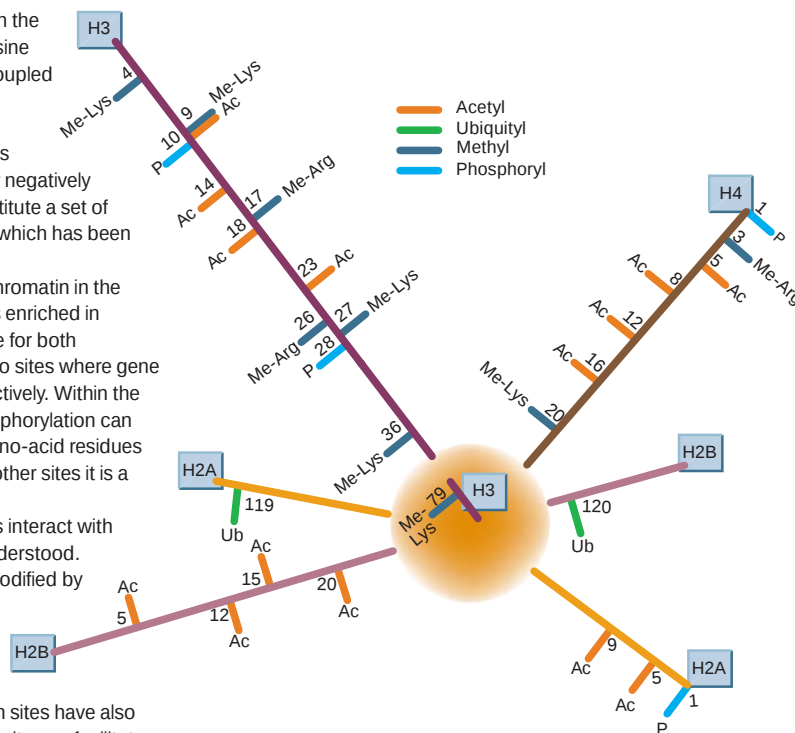
Transcription factors regulate expression by binding to specific DNA control sequences in the neighbourhood of a gene. Although some DNA sequences are accessible either as an outward-facing segment on the nucleosome surface, or in linkers between nucleosomes, most

Box 1 Histone modifications

Many amino acids of histones, particularly those in the 'tails', are chemically modified⁴⁷. These include lysine residues that may be acetylated, methylated or coupled to ubiquitin (a large polypeptide chain); arginine residues that may be methylated; and serine residues that are phosphorylated. All modifications can affect one another, and many are positively or negatively correlated with each other. Collectively, they constitute a set of markers of the local state of the genetic material, which has been called the 'histone code'⁴⁸.

Histone modification is a dynamic process. Chromatin in the neighbourhood of transcriptionally active genes is enriched in acetylated histones, and the enzymes responsible for both acetylation and deacetylation are often recruited to sites where gene expression is to be activated or repressed, respectively. Within the nucleus, local states of both acetylation and phosphorylation can change rapidly. Methylation at certain histone amino-acid residues may also be important for activation, whereas at other sites it is a signal for inactivation.

Many (perhaps all) of the histone modifications interact with each other in ways that are still not completely understood. For example, in mammals, histone H2B can be modified by ubiquitin at Lys 120 (123 in yeast), and this modification is necessary for methylation at Lys 4 and Lys 79 of histone H3, reactions that are controlled by two different methylating enzymes. Influences between nearby modification sites have also been observed, such that phosphorylation at one site can facilitate acetylation at another, methylation and phosphorylation at adjacent sites may interfere with one another, and methylation and acetylation cannot occur simultaneously on the same lysine residue.



Box 1 Figure Histone modifications. Each modification is colour coded as indicated and the position of the modified amino acid labelled^{47,49–51}.

are buried inside the nucleosome. Regulatory factors must therefore seek out their specific DNA-binding sites and gain access to them. They are aided by chromatin-remodelling complexes that continually shuffle the positions of individual nucleosomes so that sites are randomly exposed for a fraction of time^{8,14}.

A number of chromatin-remodelling complexes mobilize nucleosomes, causing the histone octamers to move short distances along the DNA⁸. Each complex carries a protein with ATPase activity, which provides the necessary energy. Many of these complexes are members of the so-called SWI/SNF family, which includes SWI/SNF in budding yeast and human, RSC in yeast, and Brahma in *Drosophila*. They have similar helicase-motif subunits, but varying co-factors within the complex. Another SWI/SNF subfamily is based on the helicase-domain protein ISWI, which combines with other proteins to form the complexes NURF, CHRAC and ACF in *Drosophila*, and RSF in humans. A third subfamily is based on the helicase motif protein Mi-2.

Remodelling complexes differ in the mechanisms by which they disrupt nucleosome structure, and they are associated with co-factors that allow them to interact selectively with other regulatory proteins that bind to specific DNA sequences. For example, only certain classes of transcription factors interact with the mammalian SWI/SNF remodelling complex. Thus remodelling complexes can be selective in the genes they modify, and transcription factors recruit these complexes as tools to gain access to chromatin.

Histone modification

Nucleosomes are not passive participants in this recognition process. They can accommodate chemical modifications — either on histone

'tails' that extend from the nucleosome surface, or within the body of the octamer — that serve as signals for the binding of specific proteins. A large number of modifications are already known, such as acetylation of amino acids in the histone tails, and new ones are being identified at a bewildering rate (Box 1). Many modifications are associated with distinct patterns of gene expression, DNA repair or replication, and it is likely that most or all modifications will ultimately be found to have distinct phenotypes.

In addition to histone modifications, nucleosomes can have core histones substituted by a variant, with functional consequences. Histone H2AZ, which is associated with reduced nucleosome stability, replaces H2A non-randomly at specific sites in the genome. Histone H2AX, which is distributed throughout the genome, is a target of phosphorylation accompanying repair of DNA breakage¹¹, and also seems to be involved in the *V(D)J* recombination events that lead to the assembly of immunoglobulin and T-cell-receptor genes. A histone H3 variant, H3.3, can be incorporated into chromatin in non-dividing cells, and seems to be associated with transcriptionally active genes¹⁰. Each of these histone substitutions is likely to be targeted by, and associated with, the binding of other proteins involved in gene activation; thus these proteins can be considered central to the formation of localized chromatin structures that are specific for gene activation or accessibility.

Interdependence of histone modifications

An interplay exists between histone modification and chromatin remodelling. For example, expression of a gene may require disruption of nucleosomes positioned at the promoter by a chromatin-remodelling complex before an enzyme required for histone

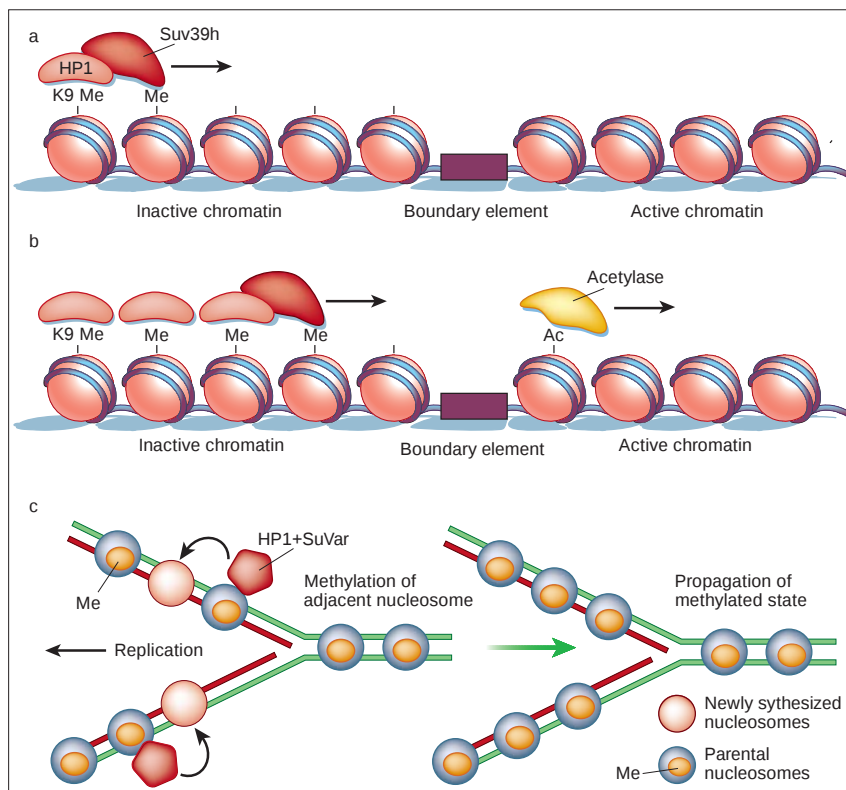


Figure 2 Propagation of inactive ('condensed') and active chromatin states (adapted from ref. 43).

a, Nucleosomes methylated at H3 Lys 9 are a mark of inactive chromatin and are bound by the heterochromatin protein HP1. HP1 in turn recruits a histone methyltransferase enzyme, Suv39h, that specifically methylates H3 Lys9, allowing methylation and HP1 binding to extend to successive nucleosomes in a self-propagating fashion^{43–45}. Some DNA sequence elements (purple rectangle) and their associated proteins may serve as barriers between different chromatin regions, perhaps by blocking the propagation of histone modifications and/or the binding of heterochromatin proteins, thus helping to establish well-defined domains⁴⁶. **b**, A similar propagation mechanism may be constructed for activation by histone acetylation (right). Here, acetylated lysines are recognized by an acetylase enzyme, resulting in acetylation of the adjacent nucleosome. **c**, A proposed model for epigenetic inheritance of methylation. During replication, parental nucleosomes carrying H3 with Lys 9 methylation (blue) are distributed randomly to both sides of the replication fork. Nucleosomes containing newly synthesized histones (pink) are deposited between the old ones, and are methylated by a mechanism similar to that described above. The daughter-cell chromatin then carries the same modification as the parent.

acetylation can be recruited¹⁵. In contrast, expression of a different gene may require that histone-acetylating enzymes and even RNA polymerase bind to the promoter prior to recruitment of the chromatin-remodelling complex¹⁶. There is no common series of steps that underlies all or even most processes of gene activation. For any given gene, however, the order of recruitment of chromatin-modifying factors may be crucial for the appropriate timing of expression.

Aside from activating gene expression, histone modifications and chromatin remodelling can also silence genes. Specific histone modifications and chromatin-remodelling complexes, such as the NuRD complex, have been implicated in silencing at some loci⁸. Even SWI/SNF complexes, which are strongly correlated with gene activation, also seem to silence a number of genes.

Specialized chromatin structures

Some regions of the genome are packaged in chromatin with distinct structural features. Three of the most studied such regions are centromeres (important for chromosomal organization during mitosis), telomeres (at the ends of chromosomes) and the inactive X chromosome in mammals. In each case, specific chromosomal structures are defined both by histones modified or substituted in specific patterns, and by the association of additional non-histone proteins or even by regulatory RNA molecules, which increasingly are implicated in chromatin organization^{17–19}.

Inactive X chromosomes in mammals are enriched for the histone variant macroH2A²⁰, which is almost three times as large as H2A itself. At vertebrate centromeres, one of the core histones, H3, is replaced by a variant, CENP-A; a similar replacement occurs in centromeres of the fruitfly *Drosophila*, indicating that this is an ancient evolutionary adaptation at centromeres. CENP-A in turn forms a complex with the centromere proteins CENP-B and -C, which mediates the formation of phased arrays of CENP-A-containing nucleosomes. In turn, additional proteins are recruited during cell division to enable the orderly separation of the two chromatids that make up each chromosome. After DNA replication, the sister chro-

matids are held together initially by a multisubunit complex called cohesin, while a second complex, condensin, helps to compact the chromosomes²¹. These complexes recognize distinct centromere structures, and a specialized nucleosome-remodelling complex associates with cohesin to help it gain access to the chromosomes²².

In the budding yeast *Saccharomyces cerevisiae*, gene silencing at the ends of chromosomes is mediated by a complex that assembles at telomeres. The complex is stabilized by the binding of the protein RAP1 to the telomere repeat sequences. Additional components, including the silent information regulator (SIR) proteins, then bind inward from the telomere ends, partly through interactions with local nucleosomes²³. One of the SIR proteins is a histone deacetylase and is thought to repress gene expression at this site. Some components of these unique complexes are evolutionarily conserved, suggesting that these unusual chromatin structures may be found in organisms other than yeast.

The silencing of genes in the vicinity of centromeres in the fission yeast *Schizosaccharomyces pombe* has been shown recently^{17–19} to depend on a set of RNA-processing enzymes involved in RNA interference, a process by which double-stranded RNA directs sequence-specific degradation of messenger RNA. One of these enzymes, Dicer, generates RNA fragments about 23 nucleotides long from transcripts of centromeric regions, which then seem in some way to be targeted back to the centromere to initiate the histone-dependent silencing mechanism. Moreover, non-coding RNA transcripts have been identified on the inactive X chromosome and elsewhere in the genome, and may have related roles at those loci²⁴.

Epigenetic inheritance

An epigenetic trait is one that is transmitted independently of the DNA sequence itself. This can occur at the level of cell division—for example, daughter cells may inherit a pattern of gene expression from parental cells (so-called cellular memory)—or at the generational level, when an offspring inherits a trait from its parents.

The classic example of epigenetic inheritance is the phenomenon of imprinting, in which the expression status of a gene depends upon

the parent from which it is derived. In mammals, for example, the *Igf2* gene (encoding insulin-like growth factor-2) is expressed only from the paternal copy of the gene, whereas the *H19* gene is expressed solely from the maternal allele. The mechanism by which this pattern of inheritance is accomplished involves (in part) DNA methylation on the paternal allele. This causes dissociation of a chromatin protein known as CTCF, which normally blocks a downstream enhancer; consequently, the enhancer is then free to activate *Igf2* expression^{25,26}.

The methylation state of an allele is linked inextricably with patterns of histone modification²⁷. Methylated CpG (guanine–cytosine) dinucleotide sites near a gene recruit specific DNA-binding proteins, which in turn recruit histone deacetylases, resulting in loss of histone acetylation and silencing of gene expression. But if histone deacetylation occurs first, it is possible to replace the acetyl group at histone H3 lysine 9 (Lys 9) with one to three methyl groups. It has been shown in turn in the fungus *Neurospora* that the ability to methylate histone Lys 9 is essential for DNA methylation²⁸, suggesting that local methylation at Lys 9 may provide a signal for methylation of the underlying DNA. Furthermore, in a different reaction pathway, maintenance of histone acetylation at promoters can lead to inhibition of DNA methylation²⁹.

Epigenetic inheritance involves the maintenance of patterns of histone modification and/or of association of chromosomal proteins correlated with specific expression states. The same mechanisms for propagating permissive or repressive chromatin structure could preserve the pattern of histone modification during replication, when old nucleosomes are distributed randomly on both sides of the fork, with the newly synthesized histones interspersed (Fig. 2).

The maintenance of repressed or activated transcription states represents an efficient mechanism for progressive cellular differentiation³⁰. In such a model, fundamental decisions regarding the turning on or off of genes or groups of genes need to be made only once. This principle is perhaps most clearly illustrated by the example of Polycomb-group (PcG)-mediated gene repression in *Drosophila*³¹. At a specific time during development, a complex of proteins, encoded by a collective of PcG genes, binds to sequences within some genes, but only in cells where the genes are silent. At subsequent stages of development, the repressed state is maintained by the PcG complex in the absence of the original negative signals. Activated expression states can be similarly maintained, again in the absence of the original transcriptional activators, by a complex of proteins encoded by genes collectively termed the trithorax group³¹. In both cases, the maintenance of gene-expression patterns is associated with specific histone modification and chromatin-remodelling activities^{32–34}.

Chromatin and nuclear self-organization

Although bacteria lack a true nucleus, a specific region of the cell, called the nucleoid, contains the chromosome, which in turn is organized into supercoiled domains or loops emanating from central nodes. The organization of the *Escherichia coli* genome into such domains is necessary to allow it to fit within the confines of the cell². Extensions of the chromosome into the cytoplasm correlate with regions that are transcriptionally active. Upon inhibition of transcription, these extensions recede to the nucleoid to give it a more even, spherical shape. The localization of genomic sequences within a bacterial cell is thus determined by their association with the transcriptional/translational apparatus.

The organization of the genome in eukaryotic nuclei, while necessarily more complex than in bacteria, seems to follow the same model as *E. coli*. Individual chromosomes largely occupy distinct 'territories' within the nucleus. Within these territories, actively transcribed genes are on surfaces of channels within subchromosomal domains³⁵ where soluble transcription factors are presumably more likely to gain access to them.

There is, however, more to the story. The eukaryotic nucleus has distinct subcompartments within which specific nuclear proteins are enriched. For example, the nucleolus, where high-level transcription

of ribosomal genes occurs, and splicing-factor compartments accumulate high local concentrations of certain proteins. In some cases there are attachment sites within the nucleus for the proteins. As a rudimentary example, one or more of the proteins associated with yeast telomeres is able to tether the telomeres in clusters to the nuclear periphery³⁶. This clustering creates a high local concentration of binding sites for the SIR silencing proteins, which in turn results in a high local concentration of these proteins, and a high occupancy of even relatively weak binding sites. The effect is to increase the extent of telomeric silencing — SIR-dependent gene silencing can be accomplished just by artificially tethering a gene to the nuclear periphery³⁷.

What organizes the formation of nuclear subdomains? Although there is evidence for a proteinaceous nuclear matrix³⁸, the example provided by yeast telomeres suggests that the chromatin fibre itself may be the organizer. Many, and probably most, chromatin-binding proteins are in continuous flux between association with chromatin and the nucleoplasm^{39,40}. Even such fundamental chromatin proteins as histone H1 have been found to bind for periods of only a few seconds, interspersed with periods of free diffusion. The notable exceptions to this rule are the core histones, the binding of which is much more stable — on the order of minutes for H2A/H2B, and hours for H3/H4. The on–off rates of proteins binding different regions of the genome may depend on the pattern of histone modifications, which in turn determines their relative enrichment in different regions of the nucleus. Thus, the genome as packaged with histones could determine the nature of nuclear subcompartments.

Future challenges

Chromatin proteins and DNA are partners in the control of the activities of the genetic material within cells. The rate-limiting step in activating gene expression typically involves alterations of chromatin structure. The chromosome is an intricately folded nucleoprotein complex with many domains, in which local chromatin structure is devoted to maintaining genes in an active or silenced configuration, to accommodating DNA replication, chromosome pairing and segregation, and to maintaining telomeric integrity. Recent results suggest strongly that in all of these cases the primary indicators of such specialization are carried on the histones. Thus, the regulatory signals that determine local properties, as well as epigenetic transmission of those properties, are likely to be on histones.

The already large catalogue of histone modifications continues to grow rapidly. Although in most cases the loss of the modification (for example, by mutating the responsible enzyme) has a detectable effect on phenotype, the function of many modifications has not yet been determined. While this will be the focus of future research, it presents significant problems because a given modification will occur at many sites in the genome, and mutations could have widespread effects, both direct and indirect. A second significant challenge arises from the potential redundancy of the 'histone code': it is possible that either of two distinct modifications could specify a single structural and functional state, or that the two modifications are always linked to one another. Significant effort will be necessary to determine the complexity of this code, that is, the number of distinct states that can be specified.

The most important immediate problem is to identify the initiating step in establishing a local chromatin state, which may also correspond to an epigenetic state. Silencing at centromeres and perhaps elsewhere seems to be initiated by small RNA transcripts from within the region to be silenced, but formation of other kinds of structures might be triggered directly by a specific histone modification. In the longer term it will be necessary to relate the reactions at individual nucleosomes to higher-order chromatin structures; this will depend in part on the development of higher-resolution methods for determining those structures, and their organization within the nucleus.

At its simplest level, chromatin should be viewed as a single entity, carrying within it the combined genetic and epigenetic codes.

Ultimately our understanding of the dynamic states of chromatin throughout the genome will be integrated with a detailed knowledge of patterns of regulation of all genes. □

doi:10.1038/nature01411

1. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp. Med.* **79**, 137–158 (1944).
2. Sandman, K., Pereira, S. L. & Reeve, J. N. Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome. *Cell Mol. Life Sci.* **54**, 1350–1364 (1998).
3. Ouzounis, C. A. & Kyriopides, N. C. Parallel origins of the nucleosome core and eukaryotic transcription from Archaea. *J. Mol. Evol.* **42**, 234–239 (1996).
4. Bailey, K. A. & Reeve, J. N. DNA repeats and archaeal nucleosome positioning. *Res. Microbiol.* **150**, 701–709 (1999).
5. Dinger, M. E., Baillie, G. J. & Musgrave, D. R. Growth phase-dependent expression and degradation of histones in the thermophilic archaeon *Thermococcus zilligii*. *Mol. Microbiol.* **36**, 876–885 (2000).
6. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).
7. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848–856 (1976).
8. Becker, P. B. & Horz, W. ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.* **71**, 247–273 (2002).
9. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–2360 (2001).
10. Ahmad, K. & Henikoff, S. Histone H3 variants specify modes of chromatin assembly. *Proc. Natl Acad. Sci. USA* **10**, 1073/pnas.172403699 (2002).
11. Redon, C. *et al.* Histone H2A variants H2AX and H2AZ. *Curr. Opin. Genet. Dev.* **12**, 162–169 (2002).
12. Smith, M. M. Centromeres and variant histones: what, where, when and why? *Curr. Opin. Cell Biol.* **14**, 279–285 (2002).
13. Wolffe, A. P. Nucleosome positioning and modification: chromatin structures that potentiate transcription. *Trends Biochem. Sci.* **19**, 240–244 (1994).
14. Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
15. Cosma, M. P., Tanaka, T. & Nasmyth, K. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* **97**, 299–311 (1999).
16. Agalioti, T. *et al.* Ordered recruitment of chromatin modifying and general transcription factors to the IFN- β promoter. *Cell* **103**, 667–678 (2000).
17. Hall, I. M. *et al.* Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237 (2002).
18. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
19. Allshire, R. RNAi and heterochromatin—a hushed-up affair. *Science* **297**, 1818–1819 (2002).
20. Chadwick, B. P. & Willard, H. F. Cell cycle-dependent localization of macroH2A in chromatin of the inactive X chromosome. *J. Cell Biol.* **157**, 1113–1123 (2002).
21. Nasmyth, K. Segregating sister genomes: the molecular biology of chromosome separation. *Science* **297**, 559–565 (2002).
22. Hakimi, M. A. *et al.* A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature* **418**, 994–998 (2002).
23. Grunstein, M. Molecular model for telomeric heterochromatin in yeast. *Curr. Opin. Cell Biol.* **9**, 383–387 (1997).
24. Kelley, R. L. & Kuroda, M. I. Noncoding RNA genes in dosage compensation and imprinting. *Cell* **103**, 9–12 (2000).
25. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).
26. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405**, 486–489 (2000).
27. Richards, E. J. & Elgin, S. C. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* **108**, 489–500 (2002).
28. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**, 556–560 (2002).
29. Mutskov, V. J., Farrell, C. M., Wade, P. A., Wolffe, A. P. & Felsenfeld, G. The barrier function of an insulator couples high histone acetylation levels with specific protection of promoter DNA from methylation. *Genes Dev.* **16**, 1540–1554 (2002).
30. Weintraub, H., Flint, S. J., Leffak, I. M., Groudine, M. & Grainger, R. M. The generation and propagation of variegated chromosome structures. *Cold Spring Harb. Symp. Quant. Biol.* **42**, 401–407 (1978).
31. Francis, N. J. & Kingston, R. E. Mechanisms of transcriptional memory. *Nature Rev. Mol. Cell Biol.* **2**, 409–421 (2001).
32. Cavalli, G. & Paro, R. Epigenetic inheritance of active chromatin after removal of the main transactivator. *Science* **286**, 955–958 (1999).
33. Sewalt, R. G. *et al.* Selective interactions between vertebrate polycomb homologs and the SUV39H1 histone lysine methyltransferase suggest that histone H3-K9 methylation contributes to chromosomal targeting of Polycomb group proteins. *Mol. Cell Biol.* **22**, 5539–5553 (2002).
34. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039–1043 (2002).
35. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* **2**, 292–301 (2001).
36. Laroche, T. *et al.* Mutation of yeast Ku genes disrupts the subnuclear organization of telomeres. *Curr. Biol.* **8**, 653–656 (1998).
37. Andrulis, E. D., Neiman, A. M., Zappulla, D. C. & Sternglanz, R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature* **394**, 592–595 (1998).
38. Hart, C. M. & Laemmli, U. K. Facilitation of chromatin dynamics by SARs. *Curr. Opin. Genet. Dev.* **8**, 519–525 (1998).
39. Misteli, T. Protein dynamics: implications for nuclear architecture and gene expression. *Science* **291**, 843–847 (2001).
40. Hager, G. L., Elbi, C. & Becker, M. Protein dynamics in the nuclear compartment. *Curr. Opin. Genet. Dev.* **12**, 137–141 (2002).
41. Alberts, B. *et al.* *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell* (Garland, New York, 1998).
42. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
43. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
44. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
45. Aagaard, L. *et al.* Functional mammalian homologues of the *Drosophila* PEV-modifier *Su(var)3-9* encode centromere-associated proteins which complex with the heterochromatin component M31. *EMBO J.* **18**, 1923–1938 (1999).
46. West, A. G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes Dev.* **16**, 271–288 (2002).
47. Goll, M. G. & Bestor, T. H. Histone modification and replacement in chromatin activation. *Genes Dev.* **16**, 1739–1742 (2002).
48. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
49. Wolffe, A. P. & Hayes, J. J. Chromatin disruption and modification. *Nucleic Acids Res.* **27**, 711–720 (1999).
50. van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**, 745–756 (2002).
51. Ng, H. H. *et al.* Lysine methylation within the globular domain of histone H3 by Dot1 is important for telomeric silencing and Sir protein association. *Genes Dev.* **16**, 1518–1527 (2002).

Acknowledgements

We thank the members of our laboratories, L. Hartwell and W. Bickmore for their advice and helpful comments. In particular, we are indebted to M. Bulger for his help in clarifying many of the issues addressed. We also acknowledge the seminal contributions to this field made by our late colleague and friend H. Weintraub, whose ideas continue to serve as a guide to our thinking.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building

4 Crinan Street

London N1 9XW, UK

Tel +44 (0) 20 7843 4961

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Senior European Sales Manager:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Hideki Watanabe

e-mail: h.watanabe@naturejpn.com

naturejobs

Alternative approaches

During the most recent glut of scientific talent, in the early to mid-1990s, young scientists who had grown weary of being serial postdocs considered their options. The choice, for many, was to leave the bench. But the big question was, to do what? The growing economy and nascent biotech boom provided some answers. Some scientists became stock-market analysts, business-development specialists or marketing experts.

Now the climate has changed. Wall Street has less value to analyse. Many biotech firms are concerned more with survival than with developing new business. And several companies have supplanted marketing with cost-cutting.

So what are the alternatives to those business-dependent careers? One would be to apply the business skills — financial analysis, intellectual property, marketing — to a different discipline, ideally one with growth potential. For example, combining expertise in materials science with knowledge of patent law could prove to be prudent — if nanotechnology ever takes off. The downside is the risk that the promise of new areas could be just as fleeting as the information-technology bubble that burst a few years ago.

Another approach is to find a more stable setting for applying 'alternative' skills. For example, doing financial analysis at a private foundation rather than a Wall Street firm, or working in the technology-transfer office of a university, rather than the business-development unit of a biotech firm. The upside is stability; the downside, scarcity.

Or perhaps, in these uncertain times, rather than looking for ways to leave the bench, now may be a good time to find ways to stay connected to it.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

The birth of
biotechnology

p456

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

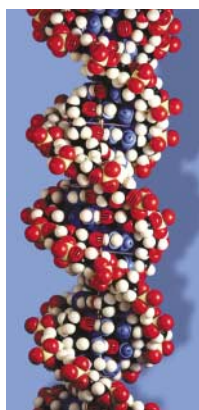
FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Learning how to manipulate DNA's double helix has fuelled job growth in biotechnology during the past 50 years, says Eugene Russo.

According to one account, biotechnology was born during a meeting at a Hawaiian delicatessen in 1972. The shop has long since been torn down, and there is no plaque to mark biotech's inception — but its legacy lives on. And the two pioneers who met there blazed distinct career paths that have become well trodden.

Stanford medical professor Stanley Cohen and biochemist Herbert Boyer from the University of California, San Francisco, were in Honolulu to attend a meeting on plasmids, the ringlets of DNA contained in bacteria. Cohen reported on the ability to introduce plasmid DNA into *Escherichia coli*, which allowed researchers to propagate and clone the plasmids in the bacteria. Boyer told the meeting about his work with a revolutionary enzyme called *EcoRI* that could cleave the double-stranded DNA molecule to produce single-stranded ends with identical termini.

Both saw the potential for combining the two discoveries into what would become genetic engineering. First, use *EcoRI* to slice both plasmid DNA and the DNA of choice. Then, with the identical DNA termini exposed, attach the DNA fragment to the plasmid DNA, and clone the whole in *E. coli*.

The two men first discussed collaboration at a deli near Waikiki Beach. Their chat over a late-night snack led to a scientific achievement that later rocked the world of science. Within a year, they had cloned DNA molecules made by splicing together DNA fragments of two different plasmids, thus creating recombinant DNA. The foundations for biotechnology were established.

Boyer and Cohen chose different paths, both affected by concerns about the safety of recombinant DNA technology (which would lead in 1975 to the Asilomar conference, where scientists, ethicists and journalists pondered the implications of genetic engineering). While Cohen stayed in academia and defended recombinant DNA technology in US congressional hearings, Boyer saw the potential for profit. In South San Francisco in 1976, Boyer and venture-

capitalist Robert Swanson set up Genentech, the world's first biotechnology company.

Pioneers at Genentech and their collaborators at the California Institute of Technology were the first to synthesize DNA in the lab. But they wanted to use *E. coli* as a factory to synthesize mammalian proteins. Proof of principle had been demonstrated earlier by Cohen and his colleagues at Stanford, when they used the bacteria to produce a functioning mouse-cell protein. The Genentech scientists eventually succeeded, producing a human hormone called somatostatin in the bacteria — and so heralded the era of commercial biotechnology. The production of insulin and growth hormone followed soon after.

In the following years, a flood of biotech firms entered the scene. Harvard professor Walter Gilbert and Phillip Sharp at the Massachusetts Institute of Technology, now both Nobel laureates, set up Biogen in Geneva in 1978. Cetus, of Emeryville, California — founded in 1971 as a 'bioengineering company' — made the push towards biotechnology and, within ten years, developed the polymerase chain reaction, which amplifies DNA. Biotech firm Amgen of Thousand Oaks, California, started up in 1980 with less than 50 employees — it now has more than 10,000 worldwide.

SCIENCE AND JUDGEMENT

Gilbert, who is now a biotech start-up specialist, says that those who have survived in the industry made astute predictions. The industry began with a focus on human proteins made in bacteria and on antibodies, he says. It then moved on to immunological treatments for cancer, and small-molecule treatments for disease. In the 1990s came a wave of neurobiology companies, followed by a wave of genomics companies. "The people who made it are the ones who guessed right," he says.

"Partly it was really good biology, but partly it was really good judgment about doing problems that were going to work," says Leroy Hood, founder of the Institute for Systems Biology in Seattle, Washington.

New technologies have fuelled the biotechnology fire. Hood and Applied Biosystems, the company he founded in 1981 in Foster City, California, came up with the automated protein synthesizer, protein sequencer, DNA synthesizer and DNA sequencer. The first three Hood calls "sophisticated plumbing" problems, as they just involved engineering a series of

Stanley Cohen, one of the founders of modern-day biotechnology.



valves to mix the correct quantities of reagents. The last was more sophisticated; it needed the integration of biology, chemistry, engineering and computer science.

Researchers also became more integrated. Many biologists joined the chemists working for big drug companies, and technological needs opened the field up to researchers with skills outside biology.

Such powerful technologies have changed the way biologists do science. 'Big science', once unique to chemistry and physics, has entered biology. Now, researchers no longer have to start their gene search with a hypothesis — with whole genomes at their disposal, they can find a gene by doing a quick database search, and then use those data in a hypothesis-driven manner for some further discovery.

A CYCLICAL INDUSTRY

Although the proliferation of big science and genomic data sparked a revolution, firms that depended too much on the human genome have faltered. Celera of Rockville, Maryland, which sequenced a draft of the human genome, has shed jobs in its effort to become a pharmaceutical company. And DoubleTwist of Oakland, California, which hoped to sell a 'superior' annotation of the human genome, went bankrupt last year.

According to Gilbert, stocks in the industry languished last year — but the industry itself did not. So small companies that are 3–5 years old are having trouble raising money at the exaggerated levels that they managed a few years ago. Gilbert says that promising new companies can still raise money — just in smaller amounts.

As a result, Steve Burrill, chief executive of Burrill & Company, a San Francisco-based biotech venture-capital firm, sees an industry with fewer jobs than in the past year, although some downsizing by older biotechs may have been offset by new start-ups. The demand for pharmacogenomics and bioinformatics expertise continues to grow, he says, along with the companies featuring them — much more than the rest of the industry. Burrill, who sees the industry as cyclical, expects both stock value and job opportunities to rebound within a year or two.

Richard Scheller, vice-president of research at Genentech, says that the company is not feeling much of a pinch, as it has a good number of products on the market. It recently hired ten new staff, and is building a new facility in South San Francisco. But even so, it is not hiring as many people as it did three years ago.

Burrill asserts that much of the restructuring in the drug industry has been good for biotech. Merging companies often shed staff, products and preclinical ideas that firms can pick up. And with pharmaceutical firms now bigger and more marketing-driven than ever, biotech companies remain important engines of innovation.

New technologies, meanwhile, continue to infuse the industry. Nanotechnology and pharmacogenomics, for example, are both areas of potential job growth. Gilbert sees 'lifestyle drugs' such as Viagra as the wave of the future — companies such as Memory Pharmaceuticals in Montvale, New Jersey, are trying to develop drugs that enhance memory and attention. Such prospects seem a far cry from the ideas raised at that Hawaiian deli more than a quarter of a century ago.

Eugene Russo is a freelance writer based in Takoma Park, Maryland.

Birth of an industry

1953 DNA structure proposed by James Watson and Francis Crick

1960 Arthur Kornberg synthesizes DNA *in vitro*

1970 Hamilton Smith and Kent Wilcox isolate the first restriction enzyme

1971 The first biological engineering company, Cetus, founded

1972 Paul Berg uses a restriction enzyme to form a hybrid circular molecule

1973 Stanley Cohen and Herbert Boyer develop DNA cloning and recombinant DNA

1975 Asilomar Conference discusses the ethics of recombinant DNA research

1976 Robert Swanson and Herbert Boyer found Genentech

- Guidelines from the US National Institutes of Health (NIH) prohibit some categories of recombinant DNA experiments

1977 Genentech clones the hormone somatostatin in bacteria, the first cloning of a protein using a synthetic recombinant gene

1978 Biogen founded, now the oldest independent biotech company

- Genentech and City of Hope National Medical Center announce the successful production of human insulin

1979 NIH recombinant DNA guidelines relaxed

- Genentech clones human growth hormone

1980 Amgen founded

- Leroy Hood and Mike Hunkapiller develop protein sequencer

- Fred Sanger and colleagues develop the shotgun method for sequencing genomes

- Bayh–Dole Act passes, encouraging technology transfer

- US Supreme Court decrees that life forms are patentable

1983 Kary Mullis invents the polymerase chain reaction

1986 US Environmental Protection Agency approves release of genetically altered tobacco, the first genetically engineered crop

1987 Susan Horvath, Mike Hunkapiller and Leroy Hood develop DNA synthesizer

1988 Philip Leder and Timothy Stewart are awarded patent for mouse breast-cancer model, the first patent given for a genetically altered animal

1993 The Biotechnology Industry Organization (BIO) formed

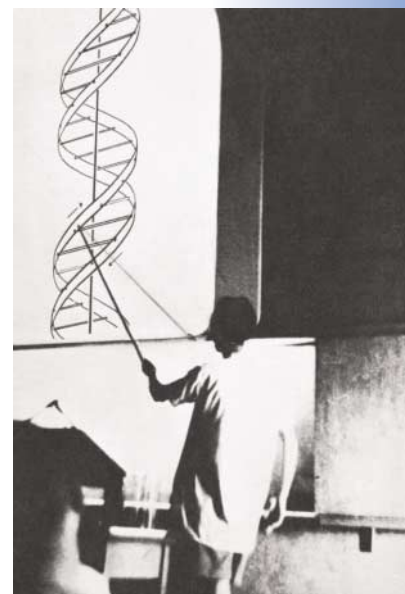
1994 US Food and Drug Administration deems the Flavr Savr, a genetically modified tomato, safe, thus bringing to market the first transgenic food product

1997 Roslin Institute scientists clone Dolly the sheep, the first mammal to be cloned successfully from an adult cell

1998 Independent teams led by James Thomson and John Gearhart are the first to grow embryonic stem cells successfully

2001 Public and private teams publish draft versions of the human genome sequence

2002 Over 200,000 jobs exist across 4,000 biotech companies worldwide, according to analysts Ernst & Young



James Watson explains the newly discovered DNA structure in a 1953 lecture



Fred Sanger: developed the shotgun sequencing method for genomes.